UTILIZING RARE AND X-LINKED VARIANTS FOR INFERENCE OF POPULATION SIZE HISTORY AND ASSOCIATION STUDIES OF COMPLEX DISEASES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Feng Gao

January 2017

UTILIZING RARE AND X-LINKED VARIANTS
FOR INFERENCE OF POPULATION SIZE HISTORY
AND ASSOCIATION STUDIES OF COMPLEX DISEASES

Feng Gao, Ph.D.
Cornell University 2017

The fast development of sequencing technologies has enabled rapid and large-scale sequencing of human genomes. This leads to the availability of an increasing number of high-quality whole-genome and exome sequencing datasets, and provides excellent opportunities for human genomic research. One common observation from these genetic datasets is an extreme excess of rare variants. One important way to utilize the information encoded in these rare variants is exploring their contribution to human complex diseases and traits. In Chapter 2, I describe our pharmacological genetic research with the goal of identifying the effects of rare genetic variants on patients' response to lipid-lowering therapies using a sequencing dataset of about 2,400 individuals. I discovered three significant associations, showing that rare variants lower the efficacy of drugs for different lipid levels. A second potential utilization of the observation of rare variants in human genomes is studying the historical scenarios that gave rise to them, specifically recent human population growth. Although many previous studies of inferring such growth from the site frequency spectrum have shown that human populations have undergone a recent epoch of fast growth in effective population size, one common limitation is that they assumed the speed of growth to take the form of exponential growth, and the ensuing models leave an excess amount of extremely rare variants. A more recent study introduced

a generalized model that allows the growth speed to be faster or slower than exponential. However, only simulation software was available for generalized models. In Chapter 3, I provide analytical expressions to accurately and efficiently evaluate the site frequency spectrum and other summary statistics under generalized models, as well as publicly available software that implements these expressions. Applying my inference framework to the data from a large-scale exome sequencing dataset, I found evidence that the recent growth of Europeans is 12% faster than exponential. Beyond autosomal variants, genetic variants on chromosome X also play a vital role in human complex diseases and quantitative traits. Compared with autosomes, chromosome X shows many unique properties, with a most obvious and important one being that males only have one copy of chromosome X. However, a vast majority of genome-wide association studies have either ignored chromosome X, or analyzed chromosome X using the same approaches for autosomal variants, potentially leading many X-linked associations to remain unrevealed. In Chapter 4, I describe XWAS, a software toolset tailored for the association analysis of chromosome X. It implements X-specific quality-control procedures as well as X-adapted single-marker and gene-based tests. I further demonstrate the usefulness of XWAS by its application to the analysis of multiple autoimmune datasets and the discovery of several new X-linked genetic associations. Some of the associations exhibit significant discrepancies in males and females, demonstrating the importance of improving association methods to account for sex bias in chromosome X.

BIOGRAPHICAL SKETCH

Feng Gao was born and raised in Shandong Province in Eastern China. In 2007, he was admitted to Fudan University, a top-tier university located in Shanghai, China, majoring in Physics. During his undergraduate study, he became interested in utilizing computer techniques to solve real-world problems. Motivated by this interest, he joined Dr. Guanghong Wei's lab in his junior year, where he performed research in computer simulations of protein folding. In his senior year, he was admitted to the Biophysics PhD Program at Cornell University with a prestigious Presidential Life Sciences Fellowship. He continued his academic journey at Cornell University right after graduation from Fudan University in 2011. During his first-year rotation at Cornell, he developed a strong interest in statistical and computational genomics. At the end of the first year, he transferred to the Computational Biology Program in order to join Dr. Alon Keinan's lab. He has been working on a variety of research problems in human population genomics and statistical/medical genetics under Dr. Keinan's mentorship in the past five years.

This document is dedicated to my parents.

ACKNOWLEDGEMENTS

First of all, I am extremely grateful to my advisor, Dr. Alon Keinan, who introduced me to the spectacular world of statistical human genetics and genomics. His knowledgeability, enthusiasm, and patience makes my PhD life enjoyable. It cannot be emphasized enough the importance of the freedom he has given me throughout my PhD study, without which I would not have been able to explore diverse aspects of human genetics and genomics, or participate in so many exciting and fruitful collaborative research. I also would like to thank past and present members of Keinan lab. I will never forget their encouragement in my life, as well as frequent and helpful discussions of my research.

Second, I am very grateful to other committee members, Andrew G. Clark, Adam C. Siepel and Haiyuan Yu, for their academic and research advice throughout my PhD study and for their kind support of different fellowship applications. My very first research experience in human genomics was actually acquired during my rotation in Dr. Clark's lab. Dr. Clark also introduced me to the valuable rotation opportunity in Dr. Keinan's lab. Without his help, I would probably not have been able to join Dr. Keinan's research group. I also benefited much from taking Dr. Siepel's probabilistic graphical model course and Dr. Yu's bioinformatics programming course. The knowledge played essential roles in my daily research.

Third, I would like to thank the Howard Hughes Medical Institute for their generous support of my last two years' PhD study with the prestigious HHMI International Student Research Fellowship. Some of my research was also supported by Dr. Keinan's awards and grants, including The Ellison Medical Foundation, The Edward Mallinckrodt, Jr. Foundation, NIH Grant R01GM108805 and NIH Grant R01HG006849.

Finally but most importantly, I would like to give my special thanks to three most cherished persons in my life. The first two, without any doubt, are my parents, who brought me to the world and raised me up, who have offered me unconditional love and support all the time, who have been supportive of every single decision I have made in my life. I also owe my heartful thanks to the third important person, my girlfriend Ting Guo, who brings sunshine and happiness to my heart every day.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The rapid reduction of human DNA sequencing costs has led to hundreds of thousands of human genomes sequenced in the past decade, enabling the availability of an increasing number of large-scale whole-genome sequencing or exome sequencing datasets (e.g., [29, 41, 44, 108, 127]) and boosting in-depth research on two major components of human genomes: rare variants, defined as variants with the less common allele present in only $< 1\%$ or $< 0.1\%$ of the sample, and X-linked variants, which are variants on chromosome X.

It is well observed that there is an extreme excess of rare variants in human genomes when the number of sequences in a sample is large. As a specific example, by analyzing the sequences of 200 genes in 14,000 individuals mostly with European ancestry, a previous study reported that more than 70% of SNPs have the minor allele presented in only one or two individuals [108]. This elevated proportion of rare variants has been suggested to be caused by recent population growth [67, 72, 118]. It has also shaped a unique left-skewed pattern in the site frequency spectrum (SFS) derived from the sequences of human populations (e.g., [29, 41, 44, 49, 108, 127]), a very important summary statistic commonly used in human genomic studies: it summarizes the proportion of variants as a function of possible allele frequency counts in the sample. In addition to the SFS, in the first section of Chapter 2, I revisited and expanded the concept of a recently presented summary statistic [67], burden of private mutation (BPM), as an alternative approach to quantify the amount of rare variants, especially singletons (SNPs with minor allele present in only one sequence among all samples), in human genomes. This summary statistic can be trans-

lated as the number of novel variants that are not present in the current sample but are expected to be ascertained with a newly sequenced individual. Based on the BPM calculated using real sequencing datasets, it was observed that after whole-genome sequencing of 4,300 individuals from a certain European population, more than 12,000 novel variants are expected with the sequencing of the next individual, which constitutes 0.5% of all heterozygous sites in the newly sequenced individual [42, 43]. This number is orders of magnitude larger than the expected value from a population that has remained a constant size throughout history (i.e., no recent growth).

This huge amount of rare variants provides us with excellent opportunities for exploration. A first potential utilization of these rare variants is elucidating their contribution to human complex diseases and traits. In the second section of Chapter 2, I describe our pharmacological genetics research that applied rare variants-tailored association approaches to the analysis of targeted sequencing data containing 5 lipid level related genes from about 2,400 patients to identify rare genetic variants that affect individuals' response to lipid-lowering therapies. We discovered three significant associations, showing that rare variants lower the efficacy of drugs for different lipid levels.

A second important use of the rare variants in human genomes is inference of recent growth of *effective* population size in human populations. Rare variants encode valuable information about very recent history of human populations because rare variants are generally younger — due to more recent mutations — than more common variants [67, 72, 118]. SFS is one of the most popular summary statistics for this inference task. By matching the SFS of a model and that derived from real sequencing data, many inference studies showed that

several human populations have undergone a recent epoch of fast growth (e.g., [29, 44, 49, 50, 108, 127]). However, all of these studies assumed that the recent growth takes the form of exponential function, which may have limited the speed of growth considering the recent faster-than-exponential growth of human *census* population size [29, 67, 117, 118]. A more recent study proposed a novel generalized model that enables the growth to be slower or faster than exponential [117, 118]. However, only simulation software were available to *simulate* genomic sequences given a generalized model [117]. The tremendous amount of time needed for simulation leaves the inference of population size history using generalized models impractical. In Chapter 3, I provide *analytical* mathematical expressions that are numerically stable and allow accurate and efficient evaluation the SFS and other summary statistics under generalized models, which were further implement in a publicly available software for population size history inference. This new software exhibits a huge speed advantage over simulation-based approaches. Applying our inference framework to the SFS of 4,300 individuals derived from a large-scale sequencing dataset [41, 127], we found strong evidence that European population has undergone a recent epoch of explosive growth that is about 12% faster than exponential.

Another substantial component of human genomes is the X chromosome (X). Previous studies have suggested that X may play an important role in sex-specific diseases such as autoimmune diseases [9, 85, 109, 115]. However, there are many dissimilarities between autosomes and X. One most important fact is, unlike autosomes, males have only one copy of X while females have two. This suggests that improved models, methods and analysis software over traditional ones designed for the association analysis of autosomal variants are needed for studying the contribution of X to complex diseases and traits. However, most of

the genome-wide association studies (GWAS) to date have either excluded X or analyzed it in the same way as for the autosomes [142], which may lead to many X-linked associations remain buried. In Chapter 4, I describe XWAS (chromosome X-Wide Analysis toolSet), a software toolset specifically designed for the association analysis of X. This collaborative work with other members of Keinan lab includes X-tailored quality control procedures as well as various X-adapted single-marker and gene-based tests. As an application, we used this new software toolset to analyze multiple autoimmune datasets, and discovered several new X-linked genetic associations. We found some of the associations show significant differences in males and females, which further demonstrate the importance of improving association methods to account for sex-bias in chromosome X.

CHAPTER 2

EXCESS OF RARE VARIANTS DUE TO RECENT EUROPEAN POPULATION GROWTH WITH APPLICATION TO ASSOCIATING RESPONSES TO PHARMACOLOGICAL GENETICS OF LIPIDS

2.1 High Burden of Private Mutations Due to Recent Explosive Growth in European Population

2.1.1 Abstract

Recent studies have shown that human populations have experienced a complex demographic history, including a recent epoch of rapid population growth that led to an excess in the proportion of rare genetic variants in humans today. This excess can impact the burden of private mutations for each individual, defined here as the proportion of heterozygous variants in each newly sequenced individual that are novel compared to another large sample of sequenced individuals. We calculated the burden of private mutations predicted by different demographic models, and compared with empirical estimates based on data from the NHLBI Exome Sequencing Project and data from the Neutral Regions (NR) dataset. We observed a significant excess in the proportion of private mutations in the empirical data compared with models of demographic history without a recent epoch of population growth. Incorporating recent growth into the model provides a much improved fit to empirical observations. This phenomenon becomes more marked for larger sample sizes, e.g. extrapolating to a scenario in which 10,000 individuals from the same population have

been sequenced with perfect accuracy, still about 1 in 400 heterozygous sites (or about 6,000 variants) at the 10,001st individual are predicted to be novel, 18-times as many as predicted in the absence of recent population growth. The proportion of private mutations is additionally increased by purifying selection, which differentially affects mutations of different functional annotations. The burden of private mutations for each individual, which are singletons (i.e. appearing in a single copy) in a larger sample that includes this individual, is predicted to be greatly increased by recent population growth, as well as by purifying selection. Comparison with empirical data supports that European populations have experienced recent rapid population growth, consistent with previous studies. These results have important implications for the design and analysis of sequencing-based association studies of complex human disease as they pertain to private and very rare variants. They also imply that personalized genomics will indeed have to be very personal in accounting for the large number of private mutations.

2.1.2 Introduction

Many recent studies that sequenced large numbers of individuals have shown that human populations have experienced a complex demographic history, including a recent epoch of rapid growth in effective population size, although estimates have varied greatly among studies [29, 41, 44, 49, 50, 108, 127]. The growth of European population has recently been estimated to be exponential with a rate of 2-5% per-generation increase in population size [41, 44, 127]. This recent growth has resulted in an excess of rare single nucleotide variants (SNVs), commonly defined as those with a minor allele (the less common of the two al-

leles) frequency (MAF) of less than 0.5% (or 1%) in a sample of individuals from the same population (e.g. [45, 108]). The proportion of singletons (SNVs with only one copy in the entire sample) is especially elevated due to this recent rapid growth [41, 44, 67, 108, 127]. Consequently, the corresponding site frequency spectrum (SFS), a summary statistic that indicates the proportion of variants of each possible allele count in the sample, is skewed towards lower allele counts (e.g. Figure 2.1).

A predicted consequence of the skew in the SFS due to population growth is an increase in the burden of private mutations for each individual. We recently defined this quantity as the proportion of heterozygous positions in each newly sequenced individual that are novel, i.e., completely absent from a previously sequenced sample from the same population [67]. In that previous paper, we observed this burden to be higher in samples from populations of European and East Asian descent than is predicted by previously estimated demographic models that do not include an epoch of recent population growth [67]. However, empirical estimates in that paper were based on a small sample size of less than 100 individuals, while the contribution of recent rapid growth is expected to be more pronounced for larger sample sizes [29, 44, 49, 50, 67, 108, 127].

Here, we set out to (1) empirically estimate the burden of private mutations from large samples of individuals of European ancestry, (2) compare these estimates with predictions of previously proposed demographic models with and without a recent epoch of exponential growth [44, 68], and (3) contrast SNVs of different functions that are expected to have undergone different selective effects. As purifying, negative selection on deleterious SNVs skews the SFS towards rare variants [20, 38, 108, 71, 127], it can interact with the effect of recent

population growth in increasing the burden of private SNVs, and differently so for different functional categories. With the rapidly decreasing cost of sequencing, more and more high-quality sequencing data sets of large sample sizes and improved accuracy of detecting rare variants become available. This provides an excellent opportunity for a more accurate study of the burden of private mutations. In this paper, we considered two such sequencing data sets of samples from populations of European ancestry: the NHLBI Exome Sequencing Project (ESP) [41, 127] and the Neutral Regions (NR) data set of putatively neutral regions [44].

2.1.3 Methods

Datasets

Two data sets were used in this study. The NR data contains the genotypes of 493 European individuals with high homogeneity on relatively neutral SNVs of 15 genetic regions [44]. For quality purposes, all SNVs with less than 900 successful genotype counts were filtered from the analysis. The remaining 1,746 SNVs constitute 95% of all variants [44]. The summarized data of 4,300 European individuals from NHLBI Exome Sequencing Project records the minor allele count and major allele count of each SNV identified in 15,585 genes on all chromosomes (including chromosome X and Y) [41, 127]. In this analysis, we combined all of the autosomal SNVs according to the 7 categories: intergenic, intron, missense, nonsense, splice, synonymous and UTR. For quality purpose, SNVs are filtered if the average read depth is less than or equal to 20 or the successful genotype counts are less than 8,170 (95%).

Subsampling Approach

In order to compare the SFS of data with different sample sizes (including the different sample sizes across the SNVs caused by unsuccessful genotype counts in the same data set), all the observed data were subsampled to 900 chromosomes. Following the strategy used in [68], for a SNV with j minor alleles out of n successful genotype counts, the probability that it is of x minor alleles when subsampled to m chromosomes is

$$\mathbb{P}(x \text{ of } m) = \frac{1}{1+\delta(x, m-x)} \left(\frac{\binom{j}{x}\binom{n-j}{m-x}}{\binom{n}{m}} + \frac{\binom{j}{m-x}\binom{n-j}{x}}{\binom{n}{m}} \right) \tag{2.1}$$

where $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ if $a \neq b$, $x = 0, 1, 2, \ldots, \lfloor \frac{m}{2} \rfloor$ and $\binom{a}{b} \coloneqq 0$ if $a < b$.

Expected SFS and the Burden of Private Mutations for Demographic Models

The SFS of the three demographic models were calculated using exact computation [6] instead of simulations.

For a demographic model with constant population size, the burden of private mutations can be derived under standard coalescent theory [99]. For constant population size, the expected number of singletons of a folded SFS for a sample of $(n+1)$ diploid individuals is

$$\mathbb{E}[\eta_1] = \theta(1 + \frac{1}{2n+1}) \tag{2.2}$$

where $\theta = 4N\mu$. The expected number of singletons that belong to one individual is

$$\mathbb{E}[s] = \frac{1}{n+1}\mathbb{E}[\eta_1] = \frac{2\theta}{2n+1} \tag{2.3}$$

The expected number of heterozygote sites for the pair of sequences from one individual $\mathbb{E}[h] = \theta$. Thus the expected burden of private mutations is

$$\mathbb{E}[\alpha] = \frac{\mathbb{E}[s]}{\mathbb{E}[h]} = \frac{2}{2n+1} \tag{2.4}$$

For variable population size, the general solution is

$$\mathbb{E}[\alpha] = \frac{1}{n+1} \frac{\mathbb{E}[T_{2n+2,1}] + \mathbb{E}[T_{2n+2,2n+1}]}{\mathbb{E}[T_{2,1}]} \tag{2.5}$$

where $T_{p,q}$ stands for the total length of all branches in the coalescent tree which have exactly q descents out of the total number of descents p. The branch lengths are calculated by exact computation [6].

Computation of the Burden of Private Mutations Using Data Sets and Simulations

For the NR data, for each of the 493 individuals, the burden of private mutations α is directly calculated by the proportion of heterozygote sites which contain singletons using the individual genotypes. Missing genotypes were abandoned. The mean and standard deviation of α for this sample were then calculated by

$$\bar{\alpha} = \frac{1}{n} \sum_{i=1}^{n} \alpha_i; \quad s(\alpha) = \sqrt{\frac{\sum_{i=1}^{n} (\alpha_i - \bar{\alpha})^2}{n-1}} \tag{2.6}$$

where n is the sample size and equals 493 here.

For ESP data and demographic models, as the individual genotypes were not available, sequences were simulated by distributing the minor alleles of each SNV to individuals randomly and independently. Unsuccessful genotype calls (missing genotypes) were also distributed randomly to the individuals but were distributed in pairs. In other words, the genotypes of each individual at each

site either were both existent or both missing. Then α was calculated using these simulated sequences in the same way as for the NR data.

For the demographic histories from which we can only get the SFS, a similar method is applied. Namely we simulated a certain number of SNVs according to the SFS and randomly assigned the minor alleles into individual sequences. The simulated sequences were paired randomly to form the sequences of an individual and α for each individual was then calculated.

To calculate α for a smaller sample size m, m individuals were randomly chosen from the original n individuals and α was calculated using the genotypes from these m individuals with the previously stated approach.

To study the effects of limited sites, a bootstrap approach was applied. Specifically, we resampled individual SNPs with replacement 1,000 times. For each bootstrap, we calculated the average α ($\alpha_{\mathrm{b},i}$) across all individuals and these 1,000 averages were used to calculate the mean and standard deviation of the bootstrap, the latter of which is an estimate of the standard error of the sample:

$$\bar{\alpha}_{\mathrm{b}} = \frac{1}{n_{\mathrm{b}}} \sum_{i=1}^{n_{\mathrm{b}}} \alpha_{\mathrm{b},i}; \quad s_{\mathrm{b}}(\alpha) = \sqrt{\frac{\sum_{i=1}^{n_{\mathrm{b}}} (\alpha_{\mathrm{b},i} - \bar{\alpha}_{\mathrm{b}})^2}{n_{\mathrm{b}} - 1}} \tag{2.7}$$

where n_{b} is the number of bootstraps and equals 1,000 here.

2.1.4 Results

In all analyses, we contrast three different demographic models and the fit of their predictions to the NR data set [44] and to 7 functional categories of the ESP data set [68, 127]. The three demographic models are (1) a population that

has been of constant population size throughout history, (2) a model of European history that includes two population bottlenecks [68], and (3) a model of European history with two bottlenecks, a recent change in population size, followed by a recent epoch of rapid population growth [44] (Model II therein).

Comparison of Site Frequency Spectra

As the burden of private mutations is a function of the site frequency spectrum, we first contrasted the site frequency spectra between three demographic models, the NR data [44], and the ESP data [68, 127] (Figure 2.1). In order to allow comparison of the data sets with different sample sizes, as well as account for missing genotype calls for each SNV, we probabilistically subsampled all data to a sample size of 900 haploid chromosomes.

The proportion of singletons from demographic models (1) and (2) is greatly lower than that in the observed data and that predicted by model (3), where recent growth is incorporated (Figure 2.1). Among the categories of the ESP data, categories that are expected to be more functional show a higher proportion of singletons, e.g. intronic, intergenic, synonymous, and UTR SNVs have a significantly lower proportion than non-synonymous, nonsense, and splice SNVs (Figure 2.1), which is expected by the latter being more often deleterious. These results recapitulate those from the ESP [41]. The proportion of singletons in the SNVs from the NR data is lower than all categories of SNVs from ESP, which is consistent with the former being designed such that variants are very far from genes and putatively neutral [44], while the latter consists of variants in and near protein-coding genes [41, 127], which are expected to more often be targeted by purifying selection. Another factor that can contribute to this

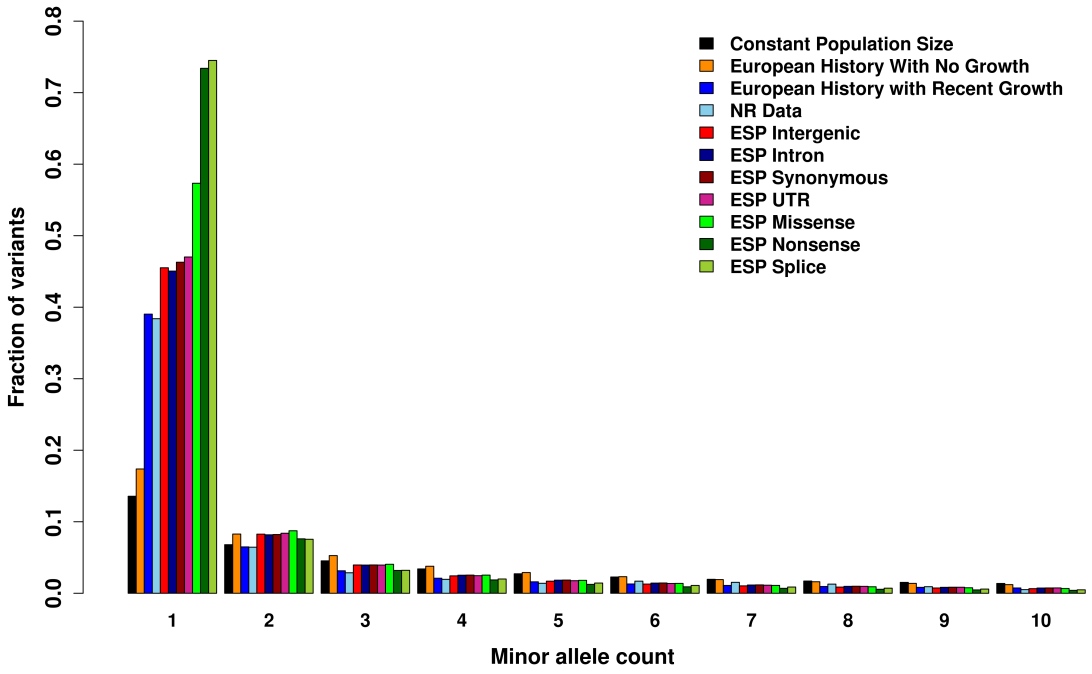

Figure 2.1: **Site frequency spectra of demographic models and data with a sample size of 900.** The SFS for 3 demographic models, the Neutral Regions (NR) data and 7 categories of the Exome Sequencing Project (ESP) data. To adjust for the different sample sizes in the two datasets, probabilistic subsampling was applied to make all sample sizes equal to 900 chromosomes. Only the first 10 minor allele count categories are shown. For each minor allele count, from left to right: constant population size, European history with 2 bottlenecks but no growth [68], European history with recent growth (Model II in [44]), the NR data, intergenic SNVs of the ESP data, intron SNVs of the ESP data, synonymous SNVs of the ESP data, UTR SNVs of the ESP data, missense SNVs of the ESP data, nonsense SNVs of the ESP data and splice SNVs of the ESP data.

difference between the NR and ESP datasets is that the former aimed to capture a sample of homogenous ancestry, which corresponds to North-Western European ancestry [44], while the latter consists of a broad sample of European Americans that exhibits a higher level of population structure [41, 127]. Increased population structure can lead to an increase in the proportion of rare variants since some of these can be due to mutations that postdate the split of the population captured by the different ancestries [44].

Comparison of the Burden of Private Mutations

The predicted burden of private mutations for each individual from all demographic models and the empirical burden observed in the different data sets and functional categories are presented in Figure 2.2. Across all sample sizes, the burden of private mutations from empirical data is significantly higher than that predicted by demographic models without growth. For example, based on the results of the NR data, when 100 individuals have been sequenced, we estimated that about 1.4% out of all heterozygous sites in the 101^{st} sequenced individual are novel, that is specific to the 101^{st} individual and completely absent from the first set of 100 individuals. While models (1) and (2) predict only 1% in this scenario, model (3) is consistent with this estimate in the NR data.

For all demographic models and observed data, as more individuals are sequenced, the burden of private mutations decreases (Figure 2.2), because increasing sample size makes it more probable that a variant has already been discovered [67]. At the same time, the effect of recent growth itself on the burden of private mutations is much more pronounced with increasing sample size. For example, for the NR data, when 492 individuals are sequenced, the estimated burden of mutation from the 493^{rd} sequenced individual is about 0.76% (Table 2.1). The estimations from models (1) and (2) are only 0.20% and 0.26%, respectively, about a third of empirical data, while model (3) matches the data well. We note that this percentage varies greatly across individuals with the relatively small number of SNVs in the NR data (Table 2.2).

When extrapolating the models to consider a scenario in which 10,000 individuals are sequenced, model (3) predicts the burden of mutations of the $10{,}001^{st}$ individual to be 0.24% (Table 2.1), 24-times and 18-times that from mod-

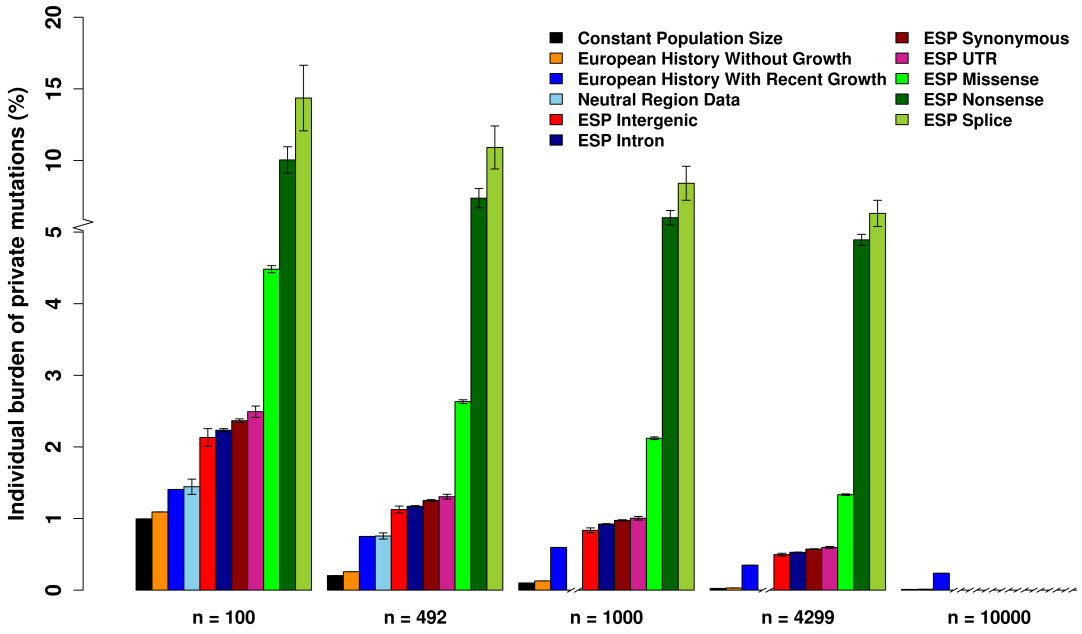

Figure 2.2: **The burden of private mutations of demographic models and empirical data.** The burden of private mutations for the same demographic models and empirical data as in Figure 2.1, using the same colors. This quantity corresponds to the percentage out of all heterozygous sites in a newly sequenced genome that are novel after n genomes have already been sequenced. Results are presented for $n = 100$, $n = 492$, $n = 1000$, $n = 4299$ and $n = 10000$. The value of 492 and 4299 are dictated by the sample size of the NR and ESP dataset, respectively. For empirical data, mean percentage across individuals is presented, together with error bars that denote $\pm$ one standard error across SNVs, estimated via bootstrapping (Methods). Double-slashes around a value of 0 on the x-axis represent instances where data for that sample size is not available in the respective dataset. Note that the range above 5% on the y-axis is rescaled. The corresponding values in this figure are shown in Table 2.1.

Table 2.1: Estimated mean and standard error of percentage of private mutations for each individual.

Group	$n = 100$	$n = 492$	$n = 1000$	$n = 4299$	$n = 10000$
Constant population size model	0.995%	0.203%	0.100%	0.023%	0.010%
European history with two bottlenecks	1.092%	0.257%	0.129%	0.031%	0.013%
European history with recent growth	1.406%	0.750%	0.596%	0.349%	0.237%
NR data	1.444% (0.106%)	0.756% (0.043%)	NA	NA	NA
ESP intergenic	2.132% (0.123%)	1.125% (0.049%)	0.835% (0.034%)	0.496% (0.019%)	NA
ESP intron	2.233% (0.022%)	1.171% (0.009%)	0.922% (0.007%)	0.528% (0.004%)	NA
ESP synonymous	2.366% (0.026%)	1.252% (0.012%)	0.974% (0.009%)	0.573% (0.005%)	NA
ESP UTR	2.492% (0.079%)	1.305% (0.034%)	1.004% (0.025%)	0.596% (0.014%)	NA
ESP missense	4.482% (0.049%)	2.632% (0.026%)	2.121% (0.019%)	1.333% (0.011%)	NA
ESP nonsense	10.04% (0.92%)	7.37% (0.68%)	6.00% (0.50%)	4.46% (0.38%)	NA
ESP splice	14.36% (2.29%)	10.91% (1.50%)	8.41% (1.19%)	6.31% (0.91%)	NA

Table 2.2: The mean and standard deviation of the burden of private mutations across individuals.

Group	The burden of private mutations
Constant population size model	0.208% (0.299%)
European history with two bottlenecks	0.276% (0.352%)
European history with recent growth	0.736% (0.614%)
NR data	0.758% (0.852%)

els without recent growth that predict 0.010% and 0.013% based on models (1) and (2), respectively (Table 2.1). This corresponds to almost 1 of 400 heterozygous positions, which is equivalent to about 6,000 variants genome-wide. This estimate is at least two orders of magnitude larger than the expected number of de novo mutations of each individual (e.g. [119]). Hence, we predict that thousands of novel variants will be discovered in each newly sequenced genome even after tens of thousands of genomes from exactly the same population have already been sequenced with perfect accuracy, and that these are rarely due to de novo mutations.

Another important observation is that the burden of private mutations for each individual calculated from all seven categories of the ESP data is consistently higher than that from the NR data for all sample sizes (Figure 2.2). This is consistent with the observation that the SFS of the ESP data are more left-skewed than those of the NR data, which is consistent with decreased effect of purifying selection and population structure on the latter. Comparing the different ESP categories, splice and non-sense SNVs, which are expected to most often be deleterious, have the largest burden of private mutations across all sample sizes. Similarly, the burden of all functional categories is ordered by common expectations as to how often such mutations are expected to be functional. The burden of private mutations captures a unique summary of the SFS that more clearly shows the effect of purifying selection. For example, when $n = 492$, the proportion of singletons is 46.2% for the ESP intergenic SNVs and 74.8% for the ESP splice SNVs, which is 1.6-fold. In comparison, the burden of private mutations for splice SNVs is about 9.7-fold of that for intergenic SNVs. This difference is even more pronounced when the sample size is larger, with 12.7-fold different when $n = 4299$ (Figure 2.2).

2.1.5 Discussion

Recent whole-genome sequencing data sets show that the proportion of rare variants in large samples, especially singletons, is significantly elevated compared with the prediction from the standard coalescent theory that assumes a constant population size and from previous demographic models without recent growth [29, 44, 67, 127]. Recent demographic modeling studies predict that humans have experienced a recent and rapid population growth, which explains an increased proportion of singletons and other rare variants [29, 44, 49, 50, 108, 127]. In this paper, we examined the burden of private mutations for each individual, a statistic that reflects the relationship between the relative proportions of singletons and more common variants contained in a sample, with three demographic models and two data sets under different sample sizes. We found that the burden of private mutations calculated from empirical data and estimated from demographic models with a recent growth is significantly higher than that estimated from models without recent growth across all sample sizes. The discrepancy is predicted to be much more pronounced for larger numbers of sequenced individuals. We showed that this finding is consistent with a recent epoch of population growth. Moreover, we found that the SNVs that are affected by stronger purifying selection will generally have larger burden of private mutations compared with more selectively neutral SNVs, since they will have a higher proportion of singletons.

The proportion of private mutations that we consider translates to the number of novel variants expected to be ascertained with each newly sequenced genome. Hence, our results have implications to sequencing-based association studies of complex human diseases and other sequencing studies. For instance,

we predict that even after 10,000 individuals from the exact same European population have been perfectly sequenced, still 1 in 400 heterozygous sites will be novel in each newly sequenced genome, which corresponds to discovering about 6,000 new variants. This large expectation is due to the effect of the recent rapid growth of European populations, which leads to this number being at least 18-fold that predicted in the absence of such growth. Hence, careful consideration must be given to private mutations in the design and analysis of sequencing-based association studies and in quantifying the role played by rare variants in complex human disease [19, 31, 40, 96, 98].

2.2 Rare LPL Gene Variants Attenuate Triglyceride Reduction and HDL Cholesterol Increase in Response to Fenofibric Acid Therapy in Individuals with Mixed Dyslipidemia

2.2.1 Abstract

Individuals with mixed dyslipidemia have elevated triglycerides (TG), low high-density lipoprotein cholesterol (HDL-C), and increased risk for coronary disease. Fibrate therapy is commonly used to lower TG and increase HDL-C. Common genetic variants are known to affect the response to fibrate therapy. We sought to identify rare genetic variants (frequency $\leq 1\%$) in genes involved in TG and HDL-C metabolism that affect the response to fenofibric acid (FA) therapy. Four genes with a major role in HDL-C and TG metabolism *APOA-I*, *APOC-II*, *APOC-III* and *LPL* were sequenced in 2,385 participants with mixed dyslipidemia in a randomized, double-blind, active-controlled study comparing therapy with FA alone, in combination with statins, or statin alone. Rare variants collapsing or SKAT methods were used for the analysis. Synonymous rare variants in the *LPL* gene were significantly associated with absolute HDL-C change ($P = 9 \times 10^{-4}$) and TG percent change ($P = 6.76 \times 10^{-4}$) in those treated with FA only. Participants with these rare variants had a 2 mg/dL increase in HDL-C and 39 mg/dL decrease in TG as compared to 6.2 mg/dL increase in HDL-C and 100 mg/dL decrease in TG in those without these variants. Rare variants in the *APOC-III* gene were associated with a modest 3 mg/dL less reduction in APOB ($P = 8.72 \times 10^{-4}$) in those receiving FA and statin. In individuals with mixed dyslipidemia rare synonymous variants within *LPL* gene

were associated with attenuated response to FA therapy while *APOC-III* rare variants were associated with a modest effect on APOB response to FA-statin therapy. These results should be replicated in a similar clinical trial for further confirmation.

2.2.2 Introduction

Multiple studies have shown that genetic variants can affect triglycerides (TG) and high-density lipoprotein cholesterol (HDL-C) response to fenofibrate. One of the common uses of fibrates including fenofibric acid (FA) is to lower TG and increase HDL-C in the population of mixed dyslipidemia (MD). Individuals with MD have high TG, low HDL-C, with or without high LDL-C, and are at higher risk for coronary heart disease. Understanding the effect of specific genetic variants on FA response can potentially help to predict its efficacy for the individual patient. Multiple common genetic variants have been shown to affect fibrate therapy [13, 14, 16, 23, 79, 88, 92], some of which had frequency as much as 20% in the mixed dyslipidemia population [13, 92]. However, although these genetic variants are frequent, their effect on drug response is usually modest.

Rare genetic variants, defined as variants with a frequency of 1% or less, have been previously shown to have a strong effect on lipid traits such as TG, HDL-C and low-density lipoprotein cholesterol (LDL-C) [26, 27]. We have recently shown that rare variants in the *APOAV* gene region have a significant effect on FA response to therapy [12].

By means of pathway approach we sought to examine the association of rare

genetic variants in a number of genes involved in TG and HDL-C metabolism pathways with levels of apolipoprotein (APO) AI, TG, APOC-III, and HDL-C in response to FA.

2.2.3 Methods

Study Population

Our study population included European-American participants from three separate concurrent prospective, randomized, double-blind, clinical trials that examined the efficacy of FA. A detailed description of the study design was included in [63, 64]. Individuals with TG $\geq$ 150 mg/dL, HDL-C $<$ 40 mg/dL in men or $<$ 50 mg/dL in women, and LDL-C $\geq$ 130 mg/dL were included. Study participants were randomized into three groups receiving either FA monotherapy, statin monotherapy, or statin-FA combination. Each study used a different statin, rosuvastatin, atorvastatin, or simvastatin. After a 6-week washout period, participants received a 12-week treatment. Lipid measurements were obtained at the beginning and end of the treatment period. The basic characteristics, including sex, age, body mass index (BMI), diabetes status, and the baseline levels (APOAI, APOB, APOC-III, HDL-C, and TG) of the three treatment groups are shown in Table 2.3.

Gene Selection

Four genes with pivotal roles in HDL-C and TG related pathways—*APOAI*, *APOC-II*, *APOC-III*, and *LPL*—were included. Each of these genes has signif-

Table 2.3: **Baseline level statistics for FA, Statin and FA + Statin treatment groups.** *P*-value is of the test for no difference among the three treatment groups using ANOVA *F*-test (age, BMI and all baseline levels) or χ^2 test (sex, diabetes).

Group	FA	Statin	FA + Statin	*P*-value
Number of individuals	358	1104	923	-
Male:Female	156:202	544:560	426:497	0.12
Age	54.81	55.05	55.07	0.98
Diabetes (With:Without)	77:281	250:854	204:719	0.89
BMI	31.8 ± 6.1	32.0 ± 6.2	31.9 ± 6.2	0.88
APOAI (mean $\pm$ sd; mg/dl)	143.0 ± 19.8	140.7 ± 20.2	138.9 ± 21.2	0.007
APOB (mean $\pm$ sd; mg/dl)	146.4 ± 25.8	144.026.2	142.626.2	0.07
APOC-III (mean $\pm$ sd; mg/dl)	18.6 ± 5.8	18.1 ± 6.0	18.0 ± 5.8	0.42
HDL (mean $\pm$ sd; mg/dl)	38.5 ± 6.6	38.7 ± 7.2	38.4 ± 7.1	0.56
TG (mean $\pm$ sd; mg/dl)	272.6 ± 136.6	275.3 ± 145.2	273.3 ± 139.0	0.93

icant impact on either HDL-C or TG metabolism and may be associated with extreme lipid phenotypes such as chylomicronemia, hypertriglyceridemia, hypobetalipoproteinemia, and elevated APOC-III. *PPARA*, the target gene for FA, which is a peroxisome proliferator-activated receptor—alpha (PPAR-alpha) agonist, was included as well.

Sequencing Protocol

Bidirectional sequencing was done at the Human Genome Sequencing Center at Baylor College of Medicine using intron-based, exon-specific primers. Polymerase chain reactions (PCR) were performed in 8 ul containing 10 ng of ge-

nomic DNA, 0.4 M oligonucleotide primers, and 0.7× Qiagen® PCR HotStar Taq Master Mix containing buffer and polymerase. Cycling parameters were 95—15 min, then 95—45 s, 60—45 s, and 72—45 s for 40 cycles followed by a final extension at 72 for 7 min. After thermocycling, 5 ul of a 1:15 dilution of Exo-SAP was added to each well, and reactions were incubated at 37 C for 15 min prior to inactivation at 80 for 15 min. Reactions were diluted by 0.6×, and 2 ul were combined with 5 ul of 1/64th Applied Biosystems® (AB) BigDye™ sequencing reaction mix and cycled as above for 25 cycles. Reactions were precipitated with ethanol, resuspended in 0.1 mM EDTA, and loaded on AB 3730XL sequencing instruments using the Rapid36 run module and 3xx basecaller. Single-nucleotide polymorphisms (SNPs) were identified using SNP Detector software [146].

Association Testing of Rare Variants

Baseline characteristics of the three treatment groups were compared using ANOVA F-test (continuous characteristics, including age, BMI, and all baseline levels) and χ^2 test (binary traits, including sex and diabetes status). Rare variants (frequency $\leq 1\%$) in the 5 sequenced genes were included in the analyses. Rare variants within a gene were further classified to categories such as intronic, missense, synonymous, promoter, and 5′ or 3′ untranslated region (UTR) variants. A total of 25 genes and gene categories (5 genes, 20 further split gene categories) were used in the analyses. The number of the rare variants within the 25 genes and gene categories, as well as the number of individuals that carry these rare variants, is shown in Table 2.4.

Two complementary statistical approaches were used in this study, the Se-

Table 2.4: Numbers of rare variants from the 25 gene and gene categories and the number of individuals with at least one rare variant for each gene and gene category.

Gene	Category	Number of rare variants	Number of individuals with at least one rare variant
APOAI	Whole gene	4	3
	Intron	4	3
APOC-II	Whole gene	98	91
	Intron	80	77
	Synonymous	3	3
	Missense	8	8
	Nonsense	1	1
	3′ UTR	2	2
	5′ UTR	4	4
APOC-III	Whole gene	12	11
	Intron	4	4
	Synonymous	5	5
	Missense	3	2
LPL	Whole gene	182	163
	Intron	107	96
	Synonymous	31	29
	Missense	25	25
	Nonsense	2	2
	Splice region	17	16
PPARA	Whole gene	152	116
	Intron	52	46
	Synonymous	26	23
	Missense	65	60
	Nonsense	1	1
	5′ UTR	8	8

quence Kernel Association Test (SKAT) and a simple collapsing approach. SKAT is a rare-variant association analysis method for sequencing data, which allows rare variants to influence the phenotype in different directions and with different magnitude of effect and shows better computational efficiency and statistical power over traditional collapsing methods [143]. This method was used to test the association between rare variants and lipid level changes after therapy in our study. The tested phenotypes were the absolute change in lipid levels (the lipid level after therapy subtracting the baseline lipid level before therapy) and percent change (the absolute lipid level change normalized by the baseline level before therapy) of five lipids or lipid-related proteins (APOAI, APOB, APOC-III, HDL-C and TG). For each analysis, six covariates were included: sex, age, BMI, smoking status, diabetes status, and baseline lipid level. To increase the statistical power of the analyses, each phenotype was tested in the three treatment groups separately. To further increase power and accuracy, all SNPs with missing rate $\geq 15\%$ were excluded from each analysis, and all individuals with missing phenotype or any missing covariate were also excluded from the analyses. The association test for each phenotype with the 25 genes and gene categories was considered a separate analysis, and a threshold value of 2.0×10^{-3} (corresponding to a nominal p-value of 0.05) was considered significant after Bonferroni correction for the 25 gene and gene categories tested.

Collapsing approaches, or "burden tests", are commonly used for association testing of rare variants [82, 102, 103]. In our study, this method was applied to show the dominating direction of the significant associations found by SKAT method, as SKAT allows rare variants to influence the phenotype in different directions and thus does not show the dominating direction of the association. We collapsed the rare variants by counting the number of rare variants in each gene

and gene category. Multivariate linear regression was then performed between the number of rare alleles and the tested phenotype using R. The regression coefficient obtained from this method is the magnitude of reduction (if the coefficient is negative) or increase (if the coefficient is positive) in the lipid level change on average when an individual has one more rare variant as compared with the individuals without any rare variant.

2.2.4 Results

Baseline characteristics of each of the three treatment groups are presented in Table 2.3. There was no significant difference at baseline between the three treatment groups other than baseline level of APOAI, which was slightly higher in the FA only treatment group. As expected, there were no differences in baseline HDL-C, TG, APOB, or APOC-III levels.

We examined the association of the 25 genes and gene categories with each of the 10 phenotypes (absolute lipid level change and percentage lipid level change of the five lipids or lipid proteins) in the three treatment groups. The Q-Q plots for p-values of the association studies are shown in Figure 2.3. Power calculation shows that SKAT method has enough power to detect significant associations with at least five rare variants. Three significant associations were found using SKAT method. These significant results were then further confirmed by 100,000 permutations.

The synonymous rare variants in the *LPL* gene region were found to be significantly associated with the absolute HDL-C change ($P = 9.00 \times 10^{-4}$, $P_c = 0.023$ after Bonferroni correction, $P_{\text{perm}} = 0.00196$ using permutation test)

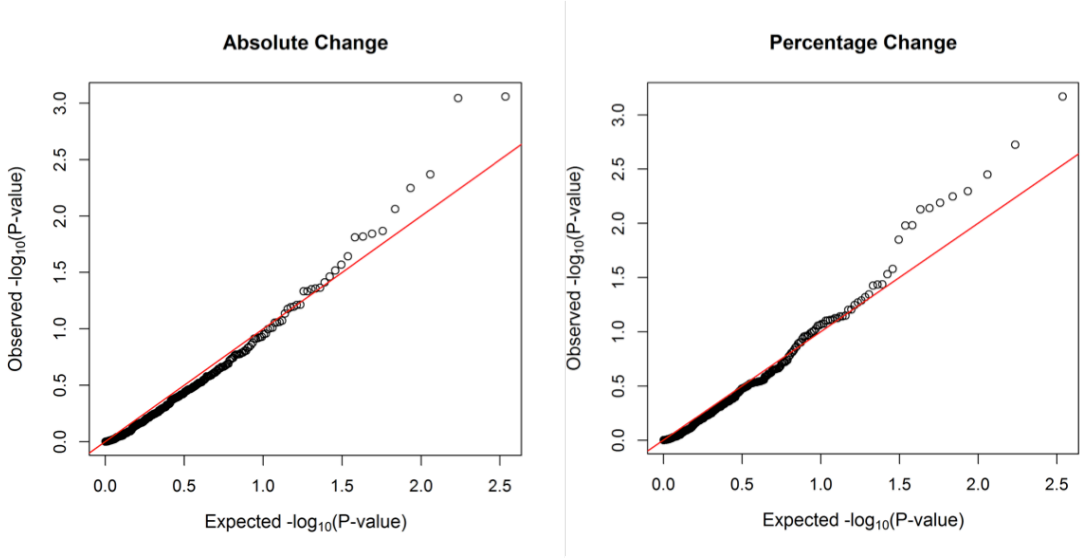

Figure 2.3: **Q-Q plots for the *P*-values from association analyses with SKAT method.** Left: the *P*-values are from the association analyses of the absolute lipid level changes of APOAI, APOB, APOC-III, HDL-C and TG; Right: the *P*-values are from the association analyses of the percentage lipid level changes of APOAI, APOB, APOC-III, HDL-C and TG.

and TG percent change ($P = 6.76 \times 10^{-4}$, $P_c = 0.017$ after Bonferroni correction, $P_{\text{perm}} = 0.00189$ using permutation test) (Table 2.5 and Table 2.6) in the FA treatment group. No association was detected with APOAI response in any of the treatment groups.

Table 2.5: Significant association between *LPL* coding synonymous gene category and absolute change in HDL-C in FA group adjusted for sex, age, BMI, smoking, diabetes and baseline HDL-C level.

Group	Number of individuals	Number of rare variants	*P*-value by SKAT
FA	318	5	9.00×10^{-4}
FA + Statin	834	11	0.50
Statin	1019	15	0.88

Participants with synonymous rare variants in the *LPL* gene region receiving FA only therapy had an attenuated increase of 2 mg/dL in HDL-C as compared

Table 2.6: Significant association between *LPL* coding synonymous gene category and percentage change in TG in FA group adjusted for sex, age, BMI, smoking, diabetes and baseline TG level.

Group	Number of individuals	Number of rare variants	*P*-value by SKAT
FA	345	5	6.76×10^{-4}
FA + Statin	899	11	0.48
Statin	1071	15	1.00

to 6.2 mg/dL increase in those without rare LPL synonymous variants (Table 2.7). The opposite pattern was observed for TG response in those receiving FA only therapy. Participants with synonymous rare variants in the *LPL* gene region had TG reduction of 39 mg/dL as compared to 100 mg/dL reduction in those without rare *LPL* synonymous variants (Table 2.8).

Table 2.7: Mean HDL-C before treatment, after treatment and change in HDL-C for individuals in FA group with and without rare variants in *LPL* coding synonymous gene category.

Group	Mean $\pm$ SE before treatment (mg/dL)	Mean $\pm$ SE after treatment (mg/dL)	Mean $\pm$ SE change (mg/dL)
Without rare variants ($n = 313$)	38.56 ± 0.38	44.78 ± 0.55	6.22 ± 0.35
With rare variants ($n = 5$)	39.88 ± 1.65	41.80 ± 6.08	2.00 ± 6.50

Combination of all of the rare variants in the *APOC-III* gene in the FA and statin combined therapy group were found to be significantly associated with the absolute change in APOB ($P = 8.72 \times 10^{-4}$, $P_c = 0.022$ after Bonferroni correction, $P_{\text{perm}} = 0.00469$ using permutation test) (Table 2.9).

Participants with rare variants in the *APOC-III* gene in the FA-statin com-

Table 2.8: Mean TG before treatment, after treatment and change in TG for individuals in FA group with and without rare variants in *LPL* coding synonymous gene category.

Group	Mean ± SE before treatment (mg/dL)	Mean ± SE after treatment (mg/dL)	Mean ± SE change (mg/dL)
Without rare variants ($n = 313$)	272.46 ± 7.82	172.27 ± 5.07	-100.19 ± 6.48
With rare variants ($n = 5$)	279.60 ± 65.31	242.20 ± 54.53	-39.40 ± 40.19

Table 2.9: Significant association between gene *APOC-III* and absolute change in APOB in FA and statin combined group adjusted for sex, age, BMI, smoking, diabetes and baseline APOB level.

Group	Number of individuals	Number of rare variants	*P*-value by SKAT
FA + Statin	887	7	8.72×10^{-4}
FA	342	0	–
Statin	1055	5	0.62

bination group had a 53 mg/dl reduction in APOB levels compared with 56 mg/dl in those without rare variants in *APOC-III*. However, there was no difference in LDL-C reduction between the groups with and without the *APOC-III* rare variants in all therapy groups (Table 2.10). The information of rare variants involved in the significant associations is listed in Table 2.11 and Table 2.12.

We used the beta coefficients from the simple collapsing method to interpret the significant results found by the SKAT method. For the FA group, the individuals with rare variants in *LPL* synonymous category tended to have lower absolute change in HDL-C and higher percentage change in TG compared with those without any rare variant (Figure 2.4). For the combination therapy group,

Table 2.10: Mean APOB before treatment, after treatment and change in APOB for individuals in FA and statin combined group with and without rare variants in *APOC-III* gene.

Group	Mean ± SE before treatment (mg/dL)	Mean ± SE after treatment (mg/dL)	Mean ± SE change (mg/dL)
Without rare variants ($n = 881$)	142.63 ± 0.88	86.14 ± 0.87	-56.49 ± 0.95
With rare variants ($n = 6$)	138.00 ± 13.57	85.33 ± 7.34	-52.67 ± 8.93

Table 2.11: Rare variants of FA group in *LPL* coding synonymous gene category.

Chromosome	Position	Minor allele	Minor allele frequency	Number of copies
8	19850095	T	2.23×10^{-4}	1
8	19857642	A	2.15×10^{-3}	1
8	19853655	C	1.31×10^{-3}	2
8	19856023	G	2.20×10^{-4}	1

Table 2.12: Rare variants of FA and statin combined group in *APOC-III* gene.

Chromosome	Position	Minor allele	Minor allele frequency	Number of copies
11	116206523	T	1.10×10^{-3}	3
11	116206603	A	2.19×10^{-4}	1
11	116206626	G	4.38×10^{-4}	1
11	116206725	G	6.59×10^{-4}	2

the individuals with rare variants in *APOC-III* gene tended to have higher absolute change in APOB compared with those without any rare variant (Figure 2.4). The corresponding information for simple collapsing method, including the Q-Q plot for the simple collapsing method is shown in Table 2.4 and Figure 2.5.

2.2.5 Discussion

In the current study we show that rare synonymous *LPL* gene variants can attenuate the effects of FA therapy on TG reduction and HDL-C increase in individuals with mixed dyslipidemia. In addition, rare *APOC-III* gene variants were associated with a modest attenuation in APOB reduction following combination therapy with statins and FA in the study population.

Rare genetic variants are fairly common but are usually unique for each individual. Overall the frequency of a specific rare variant is very low, but the probability of having some type of a unique rare variant is high. In fact, the overall population frequency of rare variants in a specific gene (i.e. total number of rare variants which are different) is higher than the frequency of many of the common SNPs in that gene. It has been previously shown that common SNPs in the *LPL* gene affect baseline triglyceride and HDL-C levels as well as response to fibrates. The additional information in this study about the effect of rare *LPL* gene variants and response to FA adds to the understanding of how *LPL* gene variants affect response to fibrate therapy.

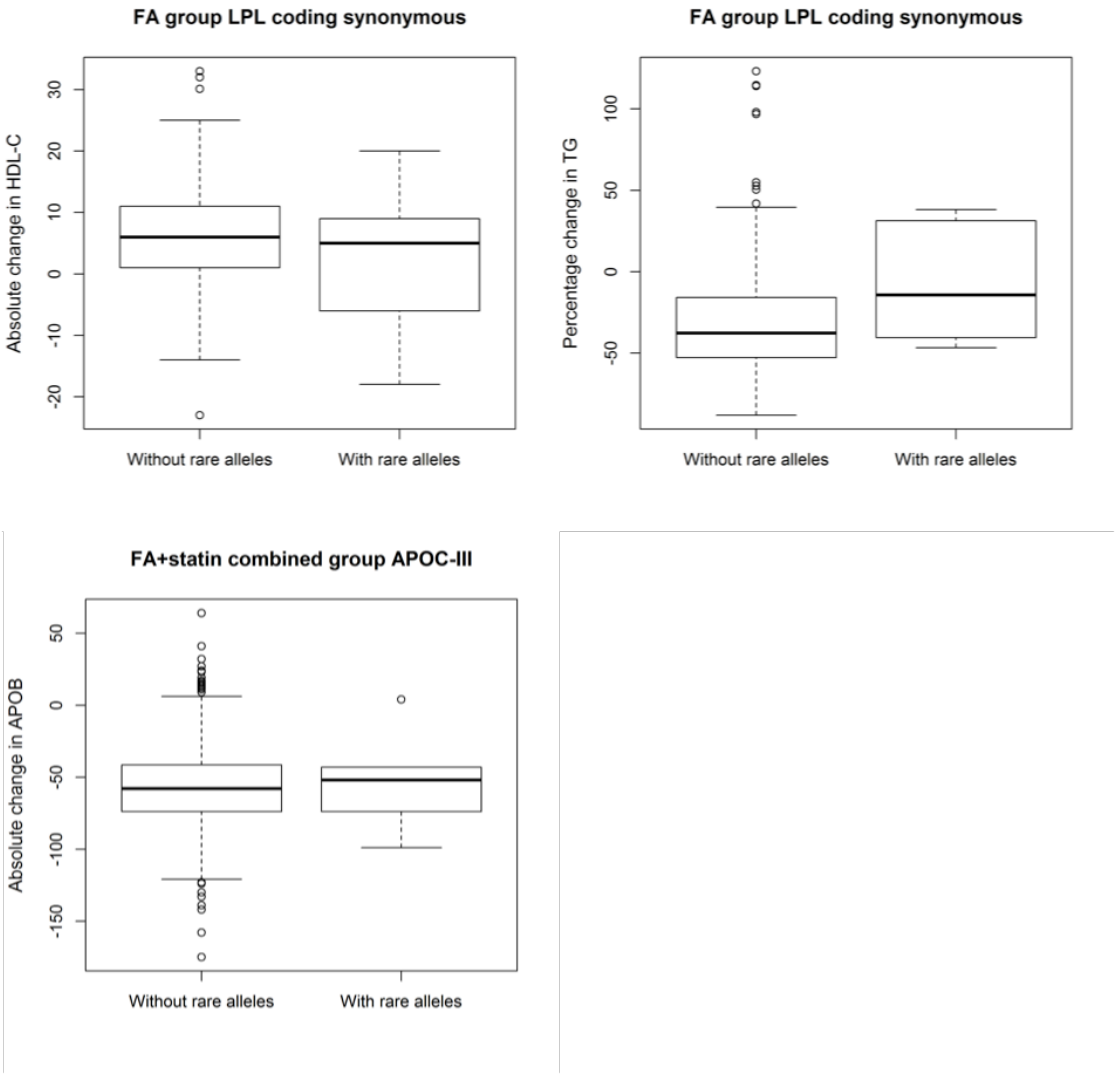

Figure 2.4: **Box plots for the three significant associations.** Left: the box plot for the association between the synonymous rare variants in the *LPL* gene region and the absolute change of HDL-C in FA group. 5 individuals have 1 rare allele in *LPL* synonymous category. Individuals with rare variants tend to have lower increase in HDL-C. Middle: the box plot for the association between the synonymous rare variants in the *LPL* gene region and the percentage change of TG in FA group. 5 individuals have 1 rare allele in *LPL* synonymous category. Individuals with rare variants tend to have lower percentage reduction in TG. Right: the box plot for the association between the variants in the *APOC-III* gene region and the absolute change of APOB. 5 individuals have 1 rare allele and 1 individual has 2 rare alleles. Individuals with rare variants tend to have lower reduction in APOB.

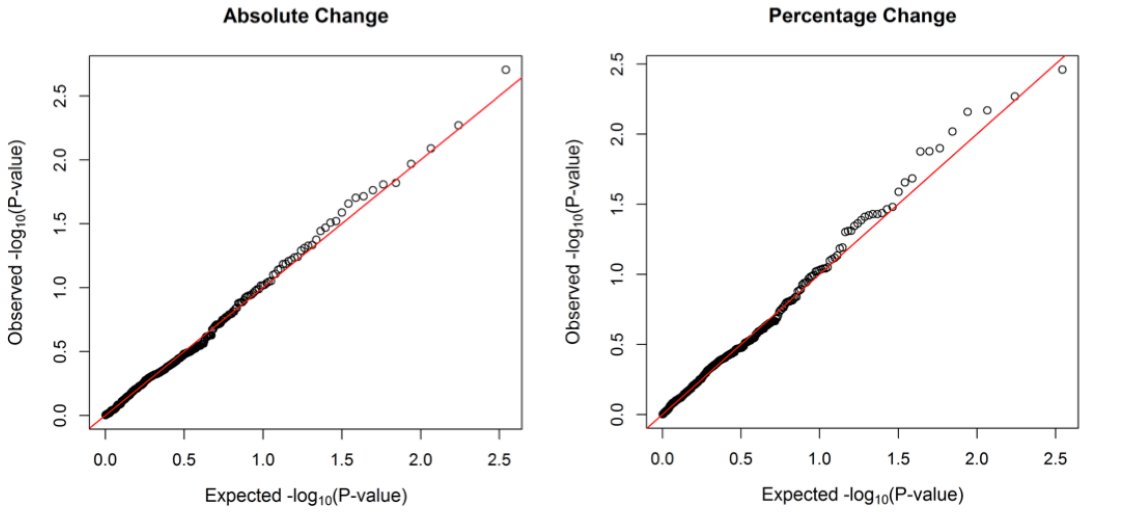

Figure 2.5: **Q-Q plots for the *P*-values from association analyses with SKAT method.** Left: the *P*-values are from the association analyses of the absolute lipid level changes of APOAI, APOB, APOC-III, HDL-C and TG; Right: the *P*-values are from the association analyses of the percentage lipid level changes of APOAI, APOB, APOC-III, HDL-C and TG.

TG reduction is thought to be a risk factor for coronary disease, as previously shown in a large Mendelian randomization study using common SNPs in the *LPL* gene region [131]. Although therapy with fibrates did not reduce coronary disease risk in the general study population in large clinical trials such as in the ACCORD LIPD trial, it did reduce coronary events in those with mixed dyslipidemia for those with absolute levels of TG $>$ 204 mg/dL and HDL-C $<$ 36 mg/dL [129]. Thus, identifying genetic factors that influence response to fibrates in the mixed dyslipidemia population has the potential to predict which patients will have higher TG reduction and larger HDL-C increase with FA therapy. This has the potential to identify which patients are most likely to derive coronary risk reduction from fibrate therapy. We suggest that rare variants in the *LPL* gene region may contribute to the coronary risk reduction effect of fibrates previously observed in the population of individuals with mixed dyslipidemia.

The LPL enzyme plays a pivotal role in TG metabolism. TG hydrolysis is facilitated by a complex interaction of LPL with various proteins such as APOC-II, APOC-III, APOAV, and others. There are well known recessive, single gene disorders that involve *APOC-II* and *LPL* genes and result in significant hypertriglyceridemia [62]. In these conditions there is little or no response to fibrates, as there is almost no residual LPL function. Common and rare SNPs may potentially have a similar but milder effect.

A common intronic SNP, rs320, was examined in a Chinese population receiving fenofibrates and was associated with a lesser TG reduction in homozygotes as compared with wild type genotypes. Although the rs328 was not expected to change the gene product, it was thought to reside in a protein binding region that indirectly effects the protein production [23]. In the current study, we show that rare synonymous *LPL* gene variants that are not expected to change protein structure can attenuate the TG reduction and HDL-C increase effects of fenofibrate.

FA is a peroxisome proliferator-activated receptor agonist that has multiple effects. It reduces TG by increasing *LPL* gene transcription and increases APOAI and HDL-C levels [135]. The rare *LPL* gene variants identified in our study may potentially affect FA response by interfering with PPAR-alpha activation of LPL transcription resulting in a smaller net TG reduction and HDL-C increase.

An additional finding was the association of the total rare variants in the *APOC-III* gene with APOB response in participants receiving the combination of statins and FA. This was a modest effect, and there is no known direct biological relationship between APOC-III and APOB. APOC-III is an important cofactor for TG hydrolysis by the LPL enzyme, and a possible explanation for

the effect of *APOC-III* rare variants on APOB response could be related to very low-density lipoprotein particle metabolism by the LPL enzyme.

Our study has limitations. The study was a randomized active-controlled trial with the most significant associations identified in the FA only group which had relatively a small sample size. This limitation was approached by using the SKAT statistical approach for which this sample size was sufficient and permutation testing that did further confirm the results. However, replication of the significant association in an independent randomized prospective clinical trial would be important to further confirm the study's results.

In conclusion, we identified rare genetic variants that affect the response to FA and its combination with statins. Our analysis suggests that synonymous variants within the *LPL* gene region may be associated with reduced response to FA in individuals with mixed dyslipidemia, which may attenuate its potential effect to reduce coronary disease.

CHAPTER 3

INFERENCE OF SUPER-EXPONENTIAL HUMAN POPULATION GROWTH VIA EFFICIENT COMPUTATION OF THE SITE FREQUENCY SPECTRUM FOR GENERALIZED MODELS

3.1 Abstract

The site frequency spectrum (SFS) and other genetic summary statistics are at the heart of many population genetic studies. Previous studies have shown that human populations have undergone a recent epoch of fast growth in effective population size. These studies assumed that growth is exponential, and the ensuing models leave an excess amount of extremely rare variants. This suggests that human populations might have experienced a recent growth with speed faster than exponential. Recent studies have introduced a generalized growth model where the growth speed can be faster or slower than exponential. However, only simulation approaches were available for obtaining summary statistics under such generalized models. In this study, we provide expressions to accurately and efficiently evaluate the SFS and other summary statistics under generalized models, which we further implement in a publicly available software. Investigating the power to infer deviation of growth from being exponential, we observed that adequate sample sizes facilitate accurate inference; e.g., a sample of 3,000 individuals with the amount of data expected from exome sequencing allows observing and accurately estimating growth with speed deviating by $\geq 10\%$ from that of exponential. Applying our inference framework to data from the NHLBI Exome Sequencing Project, we found that a model with a generalized growth epoch fits the observed SFS significantly better than the

equivalent model with exponential growth (P-value = 3.85×10^{-6}). The estimated growth speed significantly deviates from exponential (P-value $\ll 10^{-12}$), with the best-fit estimate being of growth speed 12% faster than exponential.

3.2 Introduction

Summary statistics of genetic variation play a vital role in population genetics studies, especially inference of demographic history. In particular, the site frequency spectrum (SFS) is a vital summary statistic of genetic data and is widely utilized by many demographic inference methods applied to humans and other organisms [8, 37, 50, 87, 99]. Some other demographic inference methods are based on sequential markov coalescent and utilize the most recent common ancestor (T_{MRCA}) and linkage disequilibrium patterns [53, 83, 95, 121, 122]. As another example, several studies used the average pairwise difference between chromosomes [4, 47, 51] and the SFS [69] to study the relative effective population sizes between the human X chromosome and the autosomes. The wide application of such genetic summary statistics stresses the need for their fast and accurate computation under any model of demographic history, instead of their estimations via simulations or approximations (e.g., [50, 55]).

Several recent demographic inference studies showed evidence that human populations have undergone a recent epoch of fast growth in effective population size [29, 44, 49, 50, 108, 127]. However, the above studies assumed that the growth is exponential. The observation of huge amount of extremely rare, previously unknown variants in several sequencing studies with large sample sizes [41, 108, 127] and the recent explosive growth in census population size suggests

that human population might have experienced a recent super-exponential growth, i.e. growth with speed faster than exponential [29, 67, 117, 118]. Hence, recent studies presented a new generalized growth model that extends the previous exponential growth model by allowing the growth speed to be exponential or faster/slower than exponential [117, 118]. Modeling the recent growth by this richer family of models holds the promise of a better fit to human genetic data, and can also be applicable to other organisms that experienced growth. However, only simulation approaches are currently available for evaluating such a generalized growth demographic model [117], which makes inference of demographic history computational intractable.

In this study, we first provide a set of explicit expressions for the computation of five summary statistics under a model of any number of epochs of generalized growth: (1) the time to the most recent common ancestor (T_{MRCA}), (2) the total number of segregating sites (S), (3) the site frequency spectrum (SFS), (4) the average pairwise difference between chromosomes per site (π), and (5) the burden of private mutations, BPM (α), a summary statistic that has been recently introduced as sensitive to recent growth [42, 67]. We also introduce a new software package, EGGS (Efficient computation of Generalized models' Genetic summary Statistics), that implements these expressions and facilitates fast and accurate generation of these summary statistics. We show that the numerically computed summary statistics match well with simulation results, and facilitates computation that is orders of magnitudes faster than that of simulations. By performing demographic inference on the SFS generated from simulated sequences, we then explored how many samples are needed for recovering parameters of a recent generalized growth epoch. Finally, we applied the software to investigate the nature of the recent growth in humans by inferring

demographic models using the SFS of synonymous variants of 4,300 European individuals from the NHLBI Exome Sequencing Project [41, 127].

3.3 Materials and Methods

3.3.1 Generalized Demographic Models

A demographic model $N(T)$ describes the changes of effective population size N against time T. We consider time, measured in generations, as starting from 0 at present and increasing backward in time. Furthermore, we consider the families of demographic models that are constituted by any number of epochs of generalized growth, along the lines of [7]. More formally, there exists a minimal positive integer L such that the demographic history of a population can be split into a model with $L+1$ epochs that are split by L ordered different time points $T_1, T_2, \cdots, T_L$ ($T_0 = 0 < T_1 < T_2 < \cdots < T_L < T_{L+1} = \infty$), with the k^{th} epoch starting from T_{k-1} and lasting through T_k (thus the last epoch starts at time T_L and continues into indefinite past, $T_{L+1} = \infty$). Such a history is considered as a generalized model if the population size in each epoch $N(T_{k-1} \leq T < T_k)$ can be described by the following differential equation regarding time T [117, 118]:

$$\frac{dN}{dT} = -r_k N^{b_k}, \tag{3.1}$$

where $k = 1, 2, \cdots, L+1$. Each epoch can hence capture a variety of changing patterns in effective population size. Specifically, if $r_k = 0$, this epoch is of constant population size. When $r_k \neq 0$, b_k controls the growth speed of this epoch: (1) if $b_k = 1$, the epoch is of exponential growth ($r_k > 0$) or decline ($r_k < 0$) with rate r_k; (2) if $b_k > 1$, the epoch is of faster-than-exponential (super-

exponential) growth ($r_k > 0$) or decline ($r_k < 0$); (3) if $b_k < 1$, the epoch is of slower-than-exponential (sub-exponential) growth ($r_k > 0$) or decline ($r_k < 0$). Linear growth or decline is also a special case of generalized models when $b_k = 0$. An illustration of a generalized model with 5 epochs is provided in Figure 3.1, with more detailed explanation and illustrations in Appendix A.2 and Figure A.1.

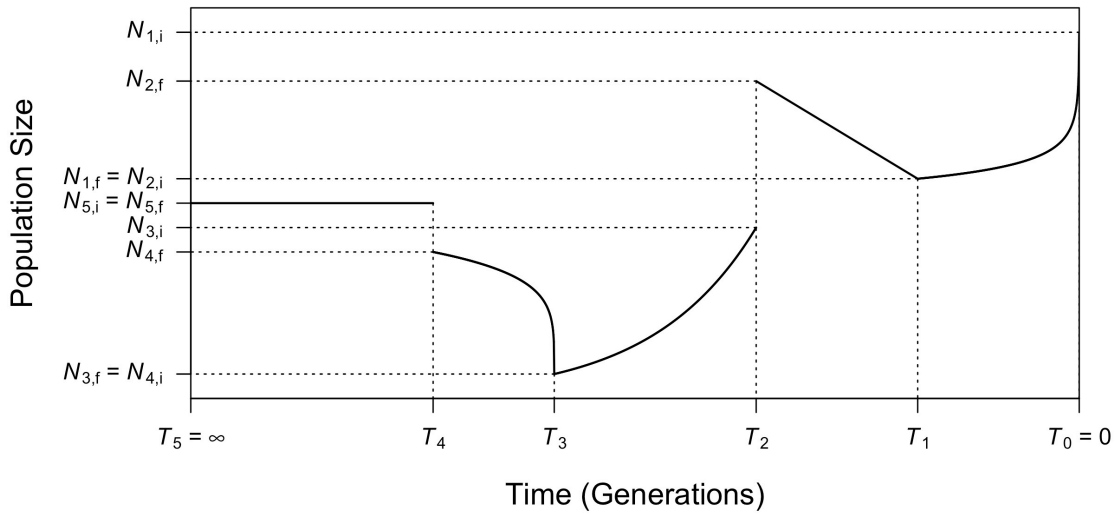

Figure 3.1: **Illustration of an example of a generalized demographic model.** This model consists of 5 epochs (starting from present): (1) Faster-than-exponential ($b > 1$) growth (looking forward in time) from $N_{1,f}$ to $N_{1,i}$ between $T_0 = 0$ and T_1; (2) Linear decline (a special case of generalized growth when $b = 0$) from $N_{2,f}$ to $N_{2,i}$ between T_1 and T_2; (3) Exponential growth (a special case of generalized growth when $b = 1$) from $N_{3,f}$ to $N_{3,i}$ between T_2 and T_3; (4) Slower-than-exponential ($b < 1$) decline from $N_{4,f}$ to $N_{4,i}$ between T_3 and T_4; (5) Constant population size (a special case of generalized growth when $r = 0$) at $N_{5,i} = N_{5,f}$ starting from T_4 and lasts indefinitely backward in time ($T_5 = \infty$). The ending population size of the previous epoch is not necessarily the beginning population size of the next epoch (e.g., $N_{2,f} \neq N_{3,i}$, $N_{4,f} \neq N_{5,i}$), corresponding to an instantaneous population size change at that time.

The solution to Equation (3.1) is [117, 118]

$$N(T) = \begin{cases} \left(N_{k,\mathrm{i}}^{1-b_k} - r_k\left(T - T_{k-1}\right)\left(1 - b_k\right)\right)^{\frac{1}{1-b_k}}, & b_k \neq 1 \\ N_{k,\mathrm{i}} \mathrm{e}^{-r_k\left(T - T_{k-1}\right)}, & b_k = 1 \end{cases} \tag{3.2}$$

where $N_{k,\mathrm{i}}$ is the initial population size of the k^{th} epoch. Each epoch k is defined by 4 parameters: the starting population size $N_{k,\mathrm{i}}$, the ending population size $N_{k,\mathrm{f}}$, the duration of the epoch $(T_k - T_{k-1})$ and the growth speed parameter b_k. The growth rate parameter, r_k is an immediate function of these parameters, $r_k = r_k\left(N_{k,\mathrm{i}}, N_{k,\mathrm{f}}, b_k, T_k - T_{k-1}\right)$, and hence does not need to be provided as an independent variable in defining the changes in effective population size during an epoch. Note that $N_{k+1,\mathrm{i}}$, the starting population size of the $(k+1)^{\mathrm{th}}$ epoch is not necessarily the same as $N_{k,\mathrm{f}}$, the ending population size of the k^{th} epoch. Specifically, if $N_{k+1,\mathrm{i}} \neq N_{k,\mathrm{f}}$, there is an instantaneous change in population size at time T_k.

3.3.2 Explicit Expressions for Summary Statistics of Demographic Models Under Arbitrary Population Size Functions

In this section, we briefly summarize the main results from previous studies that are used to evaluate the expected value of the summary statistics. Under Kingman's standard coalescent [73, 74], given a demographic model $N(T)$, the expected time of the most recent common ancestor $\mathbb{E}\left[T_{\mathrm{MRCA}}^{p}\right]$ can be calculated by [110]

$$\mathbb{E}\left[T_{\mathrm{MRCA}}^{p}\right] = \sum_{j=2}^{p} A_j^p \psi_j\,, \tag{3.3}$$

where the superscript p is the number of chromosomes (i.e. twice the sample size for diploids), ψ_j is the expected time of the first coalescent event when there are j chromosomes at present and A_j^p are constants [111, 124, 125] provided in Appendix A.3. Without loss of generality, we consider the case of diploid individuals, where there are $2N(T)$ chromosomes at any generation T, and use the notation $\mathcal{N}(T) = 2N(T)$. Then ψ_j is expressed by the following equation:

$$\psi_j = \int_0^\infty T \frac{\binom{j}{2}}{\mathcal{N}(T)} \mathrm{e}^{-\int_0^T \frac{\binom{j}{2} d\sigma}{\mathcal{N}(\sigma)}} dT = \int_0^\infty \mathrm{e}^{-\binom{j}{2}\Lambda(T)} dT, \tag{3.4}$$

where $\Lambda(T) = \int_0^T \frac{d\sigma}{\mathcal{N}(\sigma)}$.

The expected full normalized SFS $\mathbb{E}\left[\boldsymbol{\xi}^p\right] = \left(\mathbb{E}\left[\xi_1^p\right], \mathbb{E}\left[\xi_2^p\right], \ldots, \mathbb{E}\left[\xi_{p-1}^p\right]\right)$ can be computed by the following set of equations [110]:

$$\mathbb{E}\left[\xi_i^p\right] = \frac{\mathbb{E}\left[\ell_i^p\right]}{\mathbb{E}\left[\mathcal{L}^p\right]}; \mathbb{E}\left[\ell_i^p\right] = \sum_{j=2}^{p} W_{i,j}^p \psi_j; \mathbb{E}\left[\mathcal{L}^p\right] = \sum_{j=2}^{p} V_j^p \psi_j, \tag{3.5}$$

where ℓ_i^p is the length of branches in the genealogy that have i descendants ($i = 1, 2, \cdots, p-1$) and $\mathcal{L}^p = \sum_{i=1}^{p-1} \ell_i^p$ is the total length of all branches in the coalescent tree. The quantities V_j^p and $W_{i,j}^p$ are constants [110], which we provide in Appendix A.3.

Naturally, the expected number of segregating sites is given by

$$\mathbb{E}\left[S\right] = \mu_0 L \mathbb{E}\left[\mathcal{L}^p\right], \tag{3.6}$$

where μ_0 is the mutation rate per site per generation and L is the length of the locus under consideration. The average pairwise difference between chromosomes per site $\mathbb{E}\left[\pi\right]$ can be calculated by

$$\mathbb{E}\left[\pi\right] = 2\mu_0 \mathbb{E}\left[T_{\mathrm{MRCA}}^{p=2}\right]. \tag{3.7}$$

The expected burden of private mutations α at diploid sample size of $\left(\frac{p}{2}-1\right)$, defined as the proportion of heterozygous sites in a new diploid individual that are homozygous in the previous $\left(\frac{p}{2}-1\right)$ individuals, $\mathbb{E}\left[\alpha_{\frac{p}{2}-1}\right]$ can be computed by [42]

$$\mathbb{E}\left[\alpha_{\frac{p}{2}-1}\right] = \frac{2}{p\left[1+\delta\left(1,p-1\right)\right]}\frac{\mathbb{E}\left[\ell_1^p\right]+\mathbb{E}\left[\ell_{p-1}^p\right]}{\mathbb{E}\left[\ell_1^2\right]}, \tag{3.8}$$

where $\delta(\cdot,\cdot)$ is Kronecker delta function.

The detailed description of the five summary statistics mentioned above is included in Appendix A.1.

3.3.3 Evaluation of the Expected Time to the First Coalescent Event under Generalized Models

The core of evaluating the summary statistics lies in finding feasible and numerically stable functions for calculating ψ_j, the expected time of the first coalescent event when there are j chromosomes at present. Previous studies give explicit expressions of ψ_j for a demographic model constructed by exponential and constant-size epochs [8, 110]. In this study, we give a comprehensive set of ψ_j formulas for generalized models introduced above. Define $\phi_j^k \coloneqq \int_{T_{k-1}}^{T_k} \mathrm{e}^{-\binom{j}{2}\Lambda(T)}\,dT$, then $\psi_j = \sum_{k=1}^{L+1}\phi_j^k$, where $(L+1)$ is the total number of epochs. The quantity ϕ_j^k can be computed by the following set of equations:

(1) If $r_k = 0$ or $b_k = 0, r_k \neq 0$:

$$\phi_j^k = \begin{cases} \frac{1}{\binom{j}{2}}\left[\mathrm{e}^{-\binom{j}{2}\Lambda(T_k)}\mathcal{N}_{k,\mathrm{f}}\log\mathcal{N}_{k,\mathrm{f}} - \mathrm{e}^{-\binom{j}{2}\Lambda(T_{k-1})}\mathcal{N}_{k,\mathrm{i}}\log\mathcal{N}_{k,\mathrm{i}}\right], & r_k+\binom{j}{2}=0 \\ \frac{1}{r_k+\binom{j}{2}}\left[\mathrm{e}^{-\binom{j}{2}\Lambda(T_{k-1})}\mathcal{N}_{k,\mathrm{i}} - \mathrm{e}^{-\binom{j}{2}\Lambda(T_k)}\mathcal{N}_{k,\mathrm{f}}\right], & r_k+\binom{j}{2}\neq 0 \end{cases}; \tag{3.9}$$

(2) If $b_k > 0, r_k > 0$ or $b_k = 1, r_k < 0$:

$$\phi_j^k = \frac{1}{\binom{j}{2}} \left[\mathcal{N}_{k,\mathrm{i}} \mathcal{U} \left(2 - \frac{1}{b_k}, \frac{\binom{j}{2}}{b_k r_k} \mathcal{N}_{k,\mathrm{i}}^{-b_k} \right) \mathrm{e}^{-\binom{j}{2}\Lambda(T_{k-1})} - \right.$$
$$\left. \mathcal{N}_{k,\mathrm{f}} \mathcal{U} \left(2 - \frac{1}{b_k}, \frac{\binom{j}{2}}{b_k r_k} \mathcal{N}_{k,\mathrm{f}}^{-b_k} \right) \mathrm{e}^{-\binom{j}{2}\Lambda(T_k)} \right] ; \tag{3.10}$$

(3) If $b_k < 0, r_k > 0$:

$$\phi_j^k = \frac{1}{\binom{j}{2}} \left[\mathcal{N}_{k,\mathrm{f}} \mathcal{M} \left(2 - \frac{1}{b_k}, \frac{\binom{j}{2}}{b_k r_k} \mathcal{N}_{k,\mathrm{f}}^{-b_k} \right) \mathrm{e}^{-\binom{j}{2}\Lambda(T_k)} - \right.$$
$$\left. \mathcal{N}_{k,\mathrm{i}} \mathcal{M} \left(2 - \frac{1}{b_k}, \frac{\binom{j}{2}}{b_k r_k} \mathcal{N}_{k,\mathrm{i}}^{-b_k} \right) \mathrm{e}^{-\binom{j}{2}\Lambda(T_{k-1})} \right] \tag{3.11}$$

The expressions of function $\Lambda(T)$ are given in Appendix A.5. The function $\mathcal{U}(b,x) \coloneqq xU(1,b,x) = x\int_0^\infty \mathrm{e}^{-xt}(1+t)^{b-2}\,dt$, where $U(a,b,x)$ is the confluent hypergeometric function of the second kind [48]. The function $\mathcal{M}(b,x) \coloneqq \frac{x}{b-1}M(1,b,x) = x\int_0^1 \mathrm{e}^{xt}(1-t)^{b-2}\,dt$, where $M(a,b,x)$ is the confluent hypergeometric function of the first kind [48]. The exponential growth or decline then becomes a special case of $\mathcal{U}(b,x)$ when $b = 1, x \neq 0$,

$$\mathcal{U}(1,x) = x\mathrm{e}^x \int_1^\infty \frac{\mathrm{e}^{-t}}{t}\,dt = xe^x E_1(x), \tag{3.12}$$

where $E_1(x)$ is the Exponential Integral [48], which has been shown by previous studies [8, 110]. We could not find feasible and numerically stable closed-form formulas for ϕ_j^k when the population size decreases forward in time in a manner that is not linear or exponential (i.e. $r_k < 0$ and $b_k \notin \{0,1\}$). In these scenarios, we used Gauss-Legendre quadrature [66] for efficient numerical evaluation of relevant functions (see Appendix A.6 for detailed description).

3.3.4 Software Implementation

The above expressions are implemented in a software package, EGGS (Efficient computation of Generalized models' Genetic summary Statistics). The source code and compiled programs for Linux and Mac OS platforms are publicly available from website http://keinanlab.cb.bscb.cornell.edu. Source code was written in C++, with no external libraries needed for compilation. Additional information of implementation is included in Appendix A.7 and in the manual that accompanies the software online.

3.3.5 Demographic Models Assumed in This Study

The demographic models used in this study are based on the inferred European history presented by [44] (Figure 3.2, in black), which contains two bottlenecks [68] and a recent exponential growth epoch. Specifically, model presented in [44] inferred that European population had a constant effective population size of 10,000 (diploid) individuals before 4,720 generations ago, and went through the ancient bottleneck between 4,720 and 4,620 generations ago with a population size of 189. The population size then recovered to 10,000 diploids until 720 generations ago, from which time the recent bottleneck started with a size of 549. At 620 generations ago, the population size recovered to 5,633 individuals. The recent growth epoch started from 140.8 generations ago and led to a population size of 654,000 at present. The parameters of the original recent growth epoch were varied to incorporate generalized growth effects.

In addition to using the model mentioned above, we also applied an alternative model of ancient European history for inference. The model was first

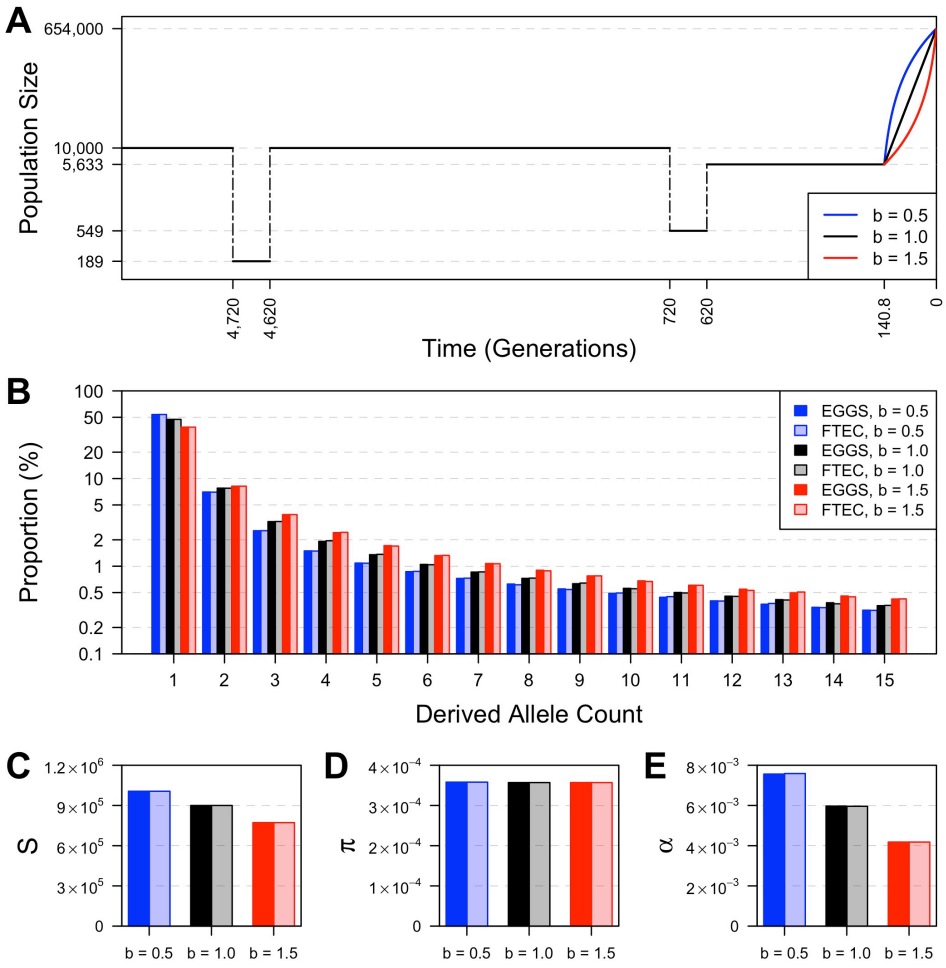

Figure 3.2: **Comparison of four summary statistics estimated by FTEC simulation and computed by EGGS.** (A) Demonstration of the demographic models considered for evaluating the accuracy of our calculations as implemented in EGGS. This two-bottleneck model has the same population size and time throughout history as in the inferred European history in [44], with the exception that we varied the growth speed parameter of the recent growth epoch to be $b = 0.5$ (sub-exponential, blue), $b = 1.0$ (exponential, black) and $b = 1.5$ (super-exponential, red). The y-axis shows effective population size of diploid individuals on log scale. (B)-(E) The comparison of the first 15 entries of the SFS (B), the total number of segregating sites (S) across all 200,000 loci (1,000 bp-long each) (C), the expected pairwise difference between chromosomes per base pair (D) and the burden of private mutation (α) as the percentage of heterozygous variants in one individual that are monomorphic in the rest of the sample of 999 individuals (E) computed numerically in EGGS (dark bars) and simulated by FTEC (light bars) for the demographic models shown in (A): $b = 0.5$, blue; $b = 1.0$, black; $b = 1.5$, red, with a sample size of 1,000 individuals (2,000 chromosomes). The y-axis in (B) is on log scale.

presented in [49] and later used in [127]. This model inferred that European population had an ancient effective population size of 7,300 diploid individuals until 6,167 generations ago, when the population size expanded to 14,474 individuals. The first bottleneck took place 2,125 generations ago, with the population size reducing to 1,861 individuals. This first bottleneck lasted until 958 generations ago, at which time a second bottleneck took place with a decreased population size of 1,032. We assumed 24 years per generation [120] to translate the year-based time presented in the original model. For compatibility with the model in [44], we considered that the population size had an instantaneous recovery after the second bottleneck lasted for 100 generations, instead of gradual recovery [44]. Figure A.2 shows the schematic representation of the model in [49].

3.3.6 Demographic Inference Framework Based on the Site Frequency Spectrum

Demographic inference in this study was based on the observed allele frequency counts from the simulated or real dataset. To determine the fitness of a model $N(T)$ to the observed data, we calculated the composite log likelihood by

$$\mathbb{L}[N] = \log \mathbb{E}\left[\boldsymbol{\xi} \mid N\right] = \boldsymbol{\mathcal{C}} \cdot \mathbb{E}\left[\boldsymbol{\xi} \mid N\right], \tag{3.13}$$

where $\boldsymbol{\mathcal{C}}$ is a vector of the observed folded allele frequency counts and $\mathbb{E}\left[\boldsymbol{\xi} \mid N\right]$ is the computed folded SFS under demographic model $N(T)$. More detailed description can be found in Appendix A.11.

To search for the maximum likelihood point over the parameter space, we applied the ECM method [101], which was previously used in the demographic

inference study by [37]. 100 ECM cycles were performed for each run of inference. We obtained 95% confidence intervals of parameter estimates via block bootstrapping of the data 200 times. Specifically, if the original data contains l loci, we randomly chose l loci from the original data with replacement in each bootstrap (see Appendix A.9 for details).

3.3.7 Processing of NHLBI Exome Sequencing Project Data for Demographic History Inference

The NHLBI Exome Sequencing Project (ESP) dataset [41, 127] contains deep sequencing of 4,300 individuals of European ancestry. An important feature of these data is the high sequencing coverage, which allows capturing very rare variants accurately. These variants constitutes the part of the SFS that is most enriched for information on recent population growth [42, 67, 127]. To reduce the effect of selection as much as possible while keeping sufficient amount of data, we chose to use the SFS calculated from synonymous single nucleotide variants (SNVs) only, as previously performed by [127]. To further improve the quality of the data, we filtered SNVs with average read depth less than or equal to 20, or with successful genotype counts smaller than 7,740 (90%), and subsampled the remaining 233,134 SNVs to 7,740 alleles, which is equivalent to 3,870 diploid individuals (Appendix A.10).

3.4 Results

3.4.1 Comparison with Simulated Results by FTEC

To validate that the expressions provided in the Methods and Materials section can correctly compute the summary statistics under generalized growth models, we compared the summary statistics calculated by our software EGGS to those simulated by the software FTEC (a coalescent simulator for modeling faster than exponential growth by [117]) under the demographic models shown in Figure 3.2(A). This model is the inferred European history in [44], except that we varied the growth speed parameter b (Equation (3.1)), which corresponds to 1 in the original model (exponential growth), to also be 0.5 (corresponding to sub-exponential growth) and 1.5 (corresponding to super-exponential growth). The sample size is fixed at 1,000 diploid individuals (2,000 chromosomes). For FTEC simulation, we used a mutation rate of 1.2×10^{-8} per base pair per generation (e.g., [75]) and simulated 200,000 independent loci, each of 1,000 base pairs.

The comparison of the SFS, S (across all 200,000 loci), π and α numerically computed by EGGS to that simulated by FTEC is shown in Figure 3.2(B)-(E). For each demographic model illustrated in Figure 2(A), the values for all summary statistics from the numerical computation by EGGS are practically identical to those from the simulation results by FTEC. However, our software EGGS exhibits a huge speed improvement over FTEC. For each model considered in Figure 3.2(A), EGGS takes less than a second to generate the results, while it takes about 5 hours for FTEC to simulate the sequences, due to the large number of independent loci required for accurate estimation (performed in Ubuntu system

with Intel Xeon CPU @ 2.67GHz). For instance, when 2,000 independent loci are simulated, which still takes about 3 minutes, the summary statistics deviate considerably from the accurate results (Figure 3.3 and Table 3.1). Furthermore, our software works well over a wide range of values of the growth parameter b, even when $b = 0$ (corresponding to linear growth or decline) or $b < 0$ (Figure 3.4), conditions that are not handled by FTEC. We note, however, that as a simulation program FTEC provides the full sequences as output and can have a wider range of applications than facilitated by the SFS and other summary statistics that EGGS calculates.

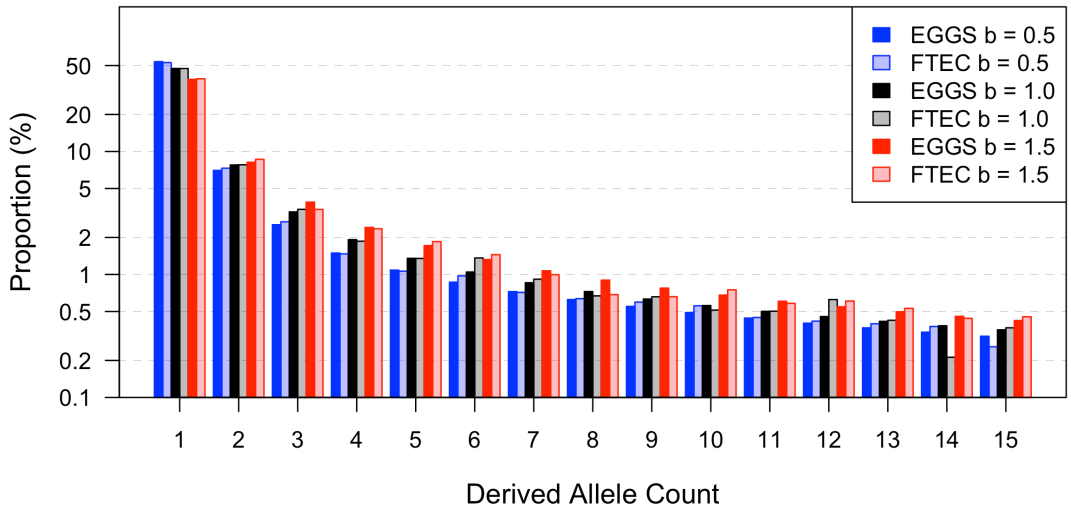

Figure 3.3: **Comparison of the first 15 entries of the SFS computed numerically in EGGS and simulated result by FTEC (light bars).** Only 2,000 loci (1,000 bp-long each) instead of 200,000 were simulated for the demographic models shown in Figure 3.2(A): $b = 0.5$, blue; $b = 1.0$, black; $b = 1.5$, red. y-axis is on log scale.

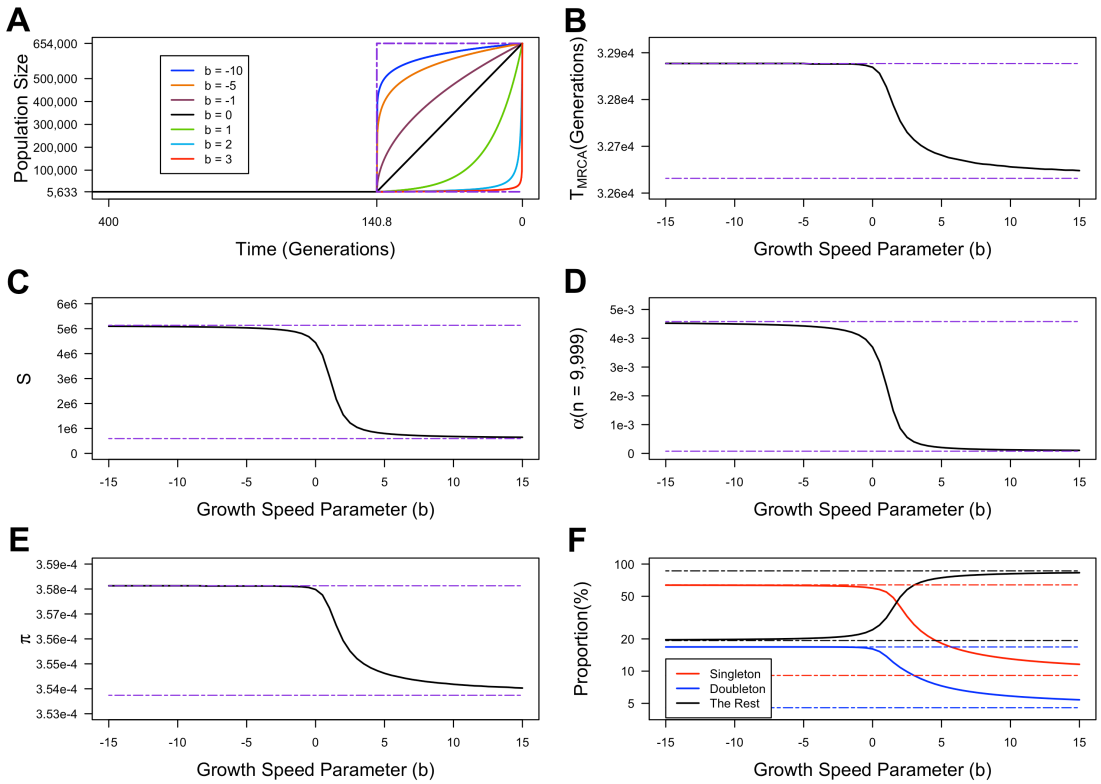

Figure 3.4: **Expected values of summary statistics generated under demographic models with a wide range of the growth speed parameter (b).** The time values and population size values are kept the same as shown in Figure 3.2(A). The growth speed parameter (b) of the recent epoch takes values from -15 to 15. The sample size is 10,000 individuals. The mutation rate per site per generation $\mu_0 = 1.2 \times 10^{-8}$. We assumed a total of 2×10^8 sites, thus the locus-based mutation rate $\mu = 2.4$. (A) The demonstration of the demographic models for several values of b. To better exhibit the difference between different values of b, only the most recent 400 generations are shown. The two dotted purple lines show the constant-size model fixed at 5,633 (corresponding to $b \to \infty$) and an instant-increase model with a sudden change from 5,633 to 654,000 at 140.8 generations ago (corresponding to $b \to -\infty$). (B)-(E) The expected value of T_{MRCA}, S, α at $n = 9{,}999$ and π respectively for b varying from -15 to 15. The two dotted purple lines correspond to the expected values for the scenarios shown by the dotted purple lines in (A). (F) The expected proportion of singletons (red), doubletons (blue) and the sum of the rest entries of the SFS for b varying from -15 to 15. The dotted lines show expected singletons (red), doubletons (blue) and the rest (black) of the SFS for the scenarios shown by the dotted purple lines in (A).

Table 3.1: **Comparison of summary statistics computed by EGGS and estimated by FTEC simulation.** Only 2,000 loci (1,000 bp-long each) were simulated for the demographic models shown in Figure 3.2(A). Presented are (i) the total number of segregating sites (S) across all 2,000 loci (1,000 bp-long each), (ii) the mean pairwise difference between chromosomes per base pair (π), and (iii) the burden of private mutation (α) as the percentage of heterozygous variants in one individual that are monomorphic in the rest of the sample of 999 individuals.

		Values of b		
		0.5	1.0	1.5
$S(10^{-4})$	EGGS	10.06	9.70	7.72
	FTEC	10.06	8.96	7.73
$\pi(10^{-4})$	EGGS	3.58	3.57	3.57
	FTEC	3.53	3.49	3.56
$\alpha(10^{-3})$	EGGS	7.56	5.97	4.18
	FTEC	7.66	6.00	4.24

3.4.2 Evaluating Inference of Generalized Growth Based on the Site Frequency Spectrum

We next set out to test the accuracy (as a function of sample size) of inferring parameters in models with generalized growth from the SFS. It has been shown in [7] that in theory, an underlying generalized growth demographic model can be uniquely identified by the ideal, perfect expected SFS with a very small sample size generated from that model (34 haploid sequences for the models shown in Figure 3.2(A)). However, the SFS is estimated in practice from a limited amount of data from each individual (even in the case of whole-genome sequencing) and as a result, the estimated SFS will fluctuate around the expected values, which limits its accuracy for inference [128]. We aim to test such inference in practice and determine the power of generalized growth detection and the sample size needed for accurately recovering the growth parameter and other parameters

of the demographic model. To be comparable with many practical applications, we considered sequence length that is about equivalent to that obtained from whole exome sequencing (Appendix A.9).

We performed inference on the SFS calculated from simulated sequences generated by FTEC. We simulated a demographic model with the same initial epochs as the model illustrated in Figure 3.2(A). Starting 620 generations ago, the simulated model included a constant population size of 10,000 until 200 generations ago, when the population starts a generalized growth epoch till present. The generalized growth epoch starts with a population size of 10,000 that grows to an extant effective population size of 1 million individuals, with the growth speed parameter b taking each of the following values: 0.4, 0.7, 0.9, 1.0, 1.1, 1.3 and 1.6. We chose these values to represent a range of super-exponential and sub-exponential growth, with emphasis on values around the exponential rate ($b = 1.0$) in order to test the detection power of generalized growth when the growth speed deviates slightly from exponential. We varied the sample size (number of diploid individuals sampled at present) to be 1,000, 2,000, 3,000, 5,000 and 10,000. The first 15 entries of the site frequency spectra for these simulated scenarios are shown in Figure 3.5. From each set of simulations, we then infer four parameters of the recent growth epoch, which can uniquely determine the epoch: 1) the growth speed parameter b; 2) the initial population size before growth N_f; 3) the ending population size after growth N_i; and 4) the onset time of growth T, which is equivalent to the growth duration since the simulated epoch ends at present.

As sample size increases, the accuracy of the point estimates generally improves and the confidence interval narrows (Figure 3.6). Specifically, when the

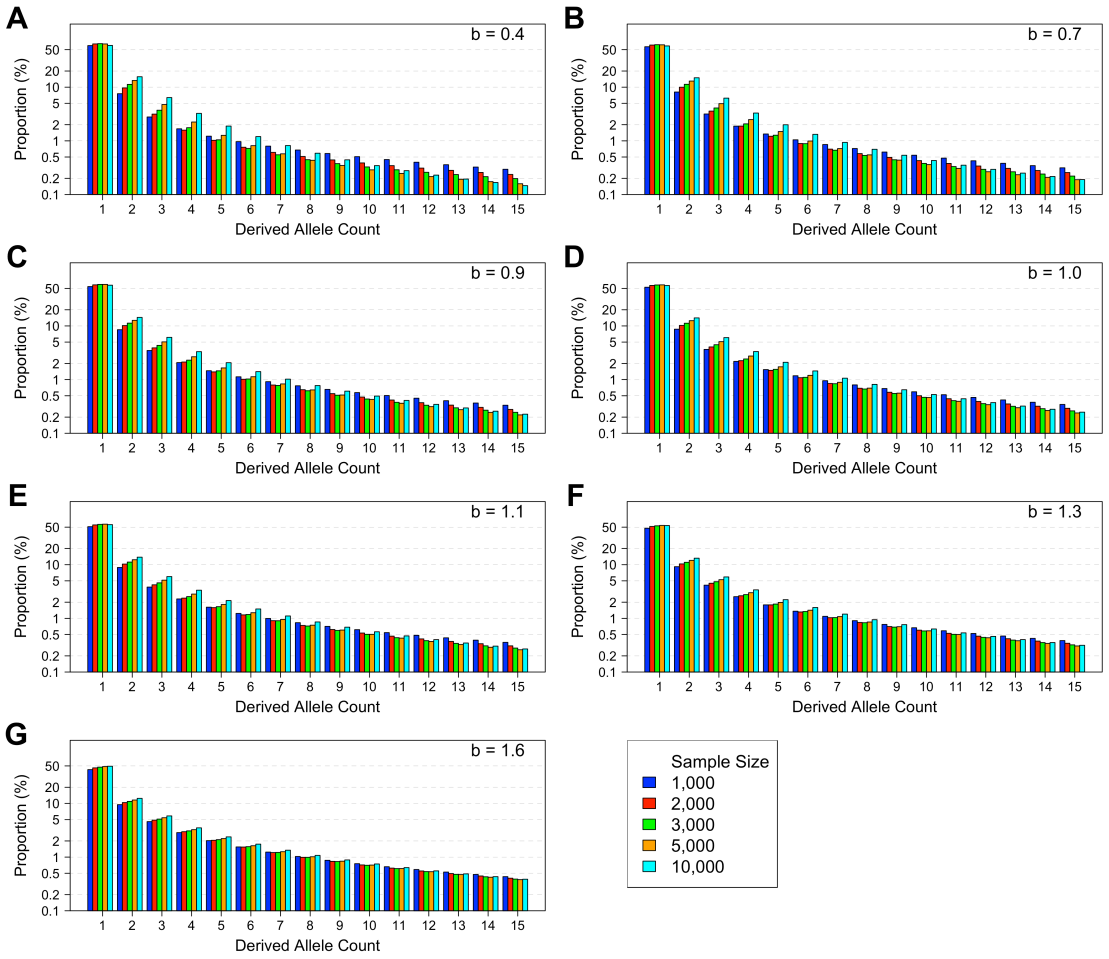

Figure 3.5: **The first 15 entries of the site frequency spectra for different simulation scenarios.** (A)-(G): corresponding to $b = 0.4$, $b = 0.7$, $b = 0.9$, $b = 1.0$, $b = 1.1$, $b = 1.3$ and $b = 1.6$ respectively for the recent generalized growth epoch, with sample size of 1,000 diploids (blue), 2,000 diploids (red), 3,000 diploids (green), 5,000 diploids (orange) and 10,000 diploids (cyan).

SFS of only 1,000 diploids is used for inference, the inference performs badly for all parameters with large confidence intervals (Figure 3.6). However, the confidence interval always includes the true simulated value. A sample size of 2,000 already exhibits acceptable performance except when the growth speed becomes large ($b = 1.3$ and 1.6). Larger sample sizes of 5,000 and 10,000 are sufficient for inferring all parameters with very tight confidence intervals. For such sample sizes, the inference even significantly distinguishes between

growth speeds ($b = 0.9$ and $b = 1.1$) that are close to exponential ($b = 1.0$) from that of an exponential, thereby concluding that a sub-exponential (0.9) or super-exponential (1.1) growth has taken place. These observations suggest that a sample size of at least 3,000 diploid individuals might be needed for inferring the parameters associated with the simulated recent generalized growth epoch, which is motivated by previous models of European demographic history. It remains to be explored how accurate the estimates are, and how their accuracy improves with sample size, across a more diverse set of models.

3.4.3 European Demographic History Inference

We next performed demographic inference on NHLBI Exome Sequencing Project (ESP) data [41, 127]. We applied our inference to these data while considering and comparing two models. Both models assume the ancient epochs before 620 generations ago to be the same as those in the model by [44] illustrated in Figure 3.2(A). We inferred the parameters only for the most recent epoch, which is of *generalized* growth in one model while limited to *exponential* growth in the other. The parameters for inference were: for both models (1) population size before growth (N_f), (2) population size after growth (N_i), (3) growth onset time (T), which is equivalent to the duration of growth; only for the *generalized* growth model (4) the growth speed parameter (b), which is fixed at $b = 1$ for the *exponential* growth model. The point estimates and 95% confidence intervals are shown in Table 3.2 and the best-fit demographic models are illustrated in Figure 3.7(AB) (also see Figure A.3, A.4 and A.5).

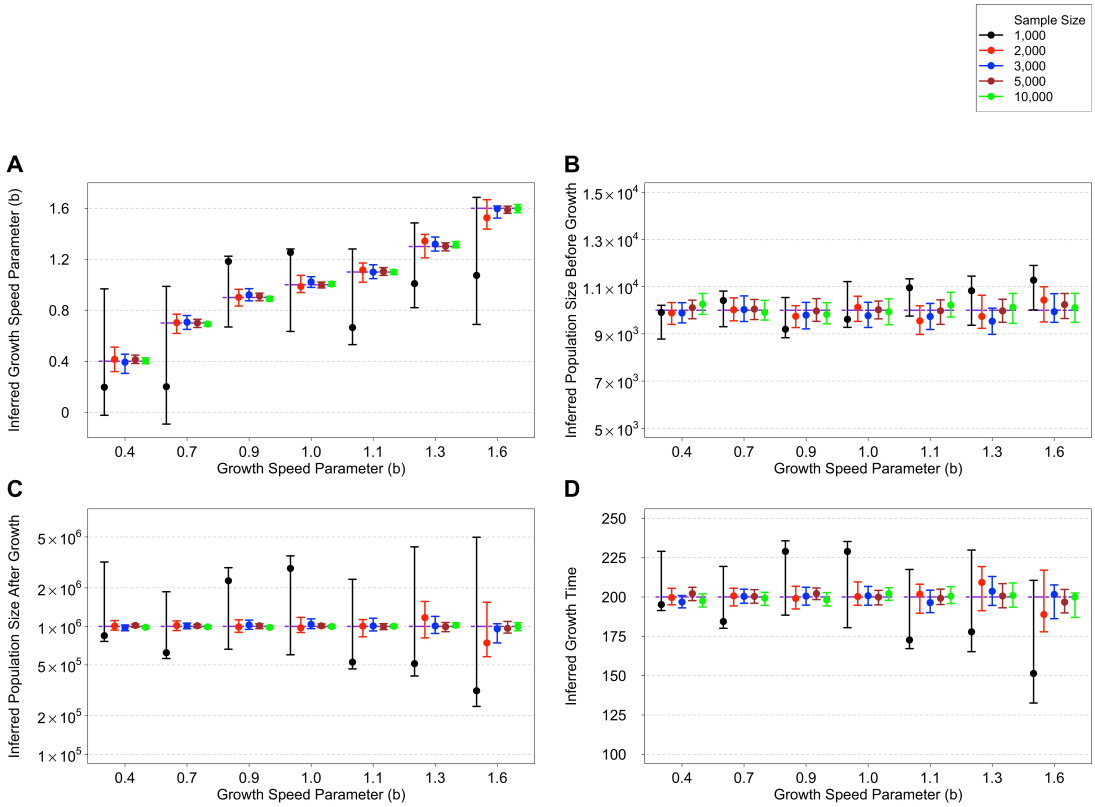

Figure 3.6: **Inference results on simulated data with a recent generalized growth epoch.** The model parameters are as follows: Growth starts 200 generations before present from an effective population size of 10,000 and ends with an effective population size of 1 million at present. The growth speed parameter b takes the following values in different simulations: 0.4, 0.7, 0.9, 1.0, 1.1, 1.3, and 1.6. Inference of these four parameters is based on the SFS estimated from a sample of individuals of one of five sizes (1,000, black; 2,000, red; 3,000, blue; 5,000, brown; and 10,000, green). The point estimates with 95% confidence interval for these parameters are grouped by the growth speed parameter b (x-axis). The dashed purple lines show the true values of the simulated model. The results are shown in the following order: (A) the inferred growth speed parameter, (B) the inferred population size before growth, (C) the inferred population size after growth, (D) the inferred growth start time. The y-axis in (C) is on log scale.

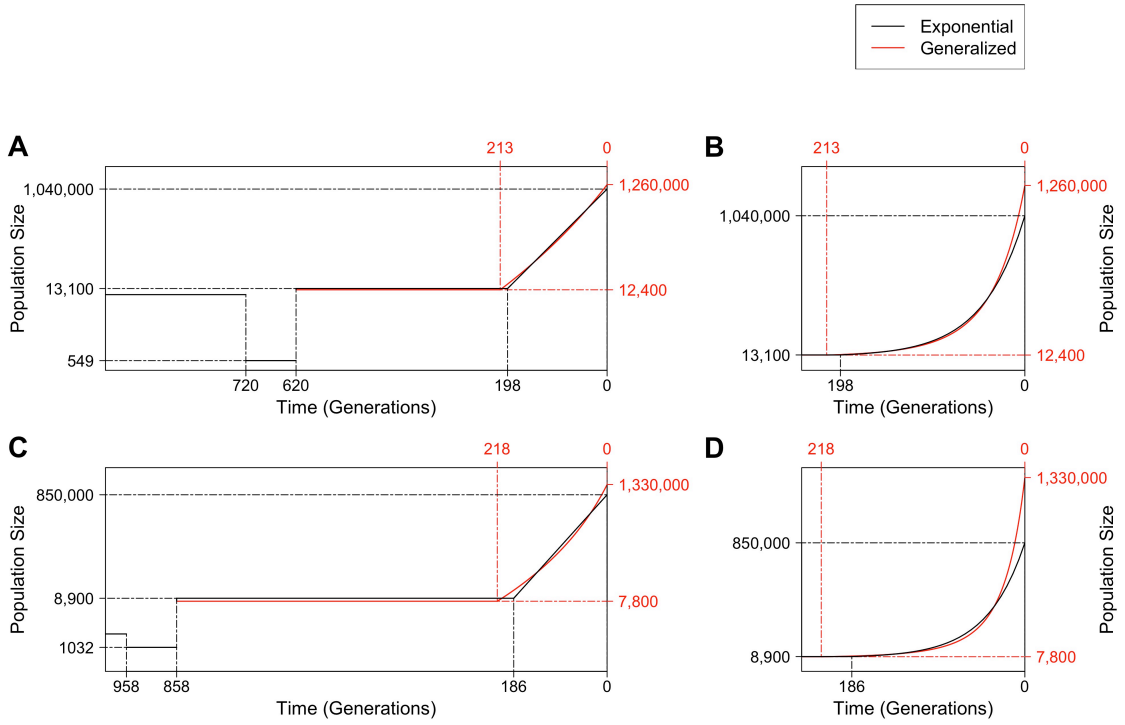

Figure 3.7: **Demographic inference results based on ESP data.** (A) Illustration of the effective population size (y-axis, on log scale) over time for the best-fit models inferred based on ESP data assuming the ancient history is the same as that in [44]. Two models are shown: One restricted to recent growth being exponential (black) and one with a generalized recent growth epoch (red). Before 620 generations ago, the model was not inferred and all parameters were set to be the same as those shown in Figure 3.2(A). Solid lines show the effective population size over time of each of the inferred models, with dashed line indicating estimated parameter values on the x-axis or y-axis. Only the most recent 1,000 generations are shown to emphasize the difference between the two models. (B) A zoom-in to the most recent 240 generations of the inferred models in (A) to emphasize the acceleration pattern of the generalized growth model, with y-axis on linear scale. (C) Similar to panel (A), except that the best-fit models presented are based on the assumption that the ancient history before 858 generations ago is fixed to that in [49] (see Figure A.2). (D) A zoom-in to the most recent 240 generations of the inferred models in (C).

Table 3.2: **Demographic inference results using ESP data for a model with a recent epoch of exponential growth and a model with a recent epoch of generalized growth.** Shown are point estimates and 95% confident intervals (in parenthesis) for the following parameters of the inferred recent growth epoch when the ancient history was assumed to be the same as that in Gazave *et al.* (2014) model and Gravel *et al.* (2011) model: population size before growth (N_f), population size after growth (N_i), time growth started in generations (T), and the growth speed parameter (b), which is fixed at $b = 1$ in the exponential growth case.

Ancient History	Growth Model	N_f (10^4)	N_i (10^6)	T	b
Gazave Model	Exponential	1.31 (1.26-1.36)	1.04 (1.00-1.07)	198 (195-202)	N/A
	Generalized	1.24 (1.18-1.30)	1.26 (1.16-1.37)	213 (206-220)	1.12 (1.07-1.15)
Gravel Model	Exponential	0.89 (0.86-0.93)	0.85 (0.82-0.88)	186 (182-190)	N/A
	Generalized	0.78 (0.74-0.83)	1.33 (1.22-1.46)	218 (211-228)	1.22 (1.18-1.26)

Although the model by [44] assumed a different ancient history before the recent growth epoch than that assumed in [127], using ESP data and assuming *exponential* growth, the inferred growth epoch is generally consistent with that obtained in their study (Figure 3.7(AB) and Table 3.2). Our study infers that recent growth started 198 (95% CI: 195-202) generations ago with an effective population size of ∼13,100 (12,600-13,600) and continued at a rate of 2.2% (2.15%-2.26%) per generation (Table 3.2), while [127] estimated that recent growth had an initial population size of ∼9,500 individuals, a duration of 204 generations and a growth rate of 2.0% per generation.

The inferred *generalized* growth model fits the data significantly better than that with *exponential* growth (P-value $= 3.85 \times 10^{-6}$ by χ^2 likelihood-ratio test with 1 degree of freedom). It estimates that growth started 213 (206-220) genera-

tions ago from an effective population size of 12,400 (11,800-13,000), both values consistent with those estimated in the *exponential* growth model. The extant effective population size following growth is estimated to be 1.26 (1.16-1.37) million. The inferred growth speed parameter $b = 1.12$ (1.07-1.15) is significantly larger than exponential speed of $b = 1$ (P-value $\ll 10^{-12}$ using one-tailed z-test), which is the main difference between the two models. $b = 1.12$ implies a growth rate acceleration pattern that is super-exponential at 12% faster than exponential through the epoch (Figure 3.7): the super-exponential growth is relatively slow around the onset time, and it keeps accelerating as time approaches present.

To test the sensitivity of the model to the assumption of ancient European history, we considered an alternate model of ancient history. We fixed the history before 858 generation ago to be that inferred by [49] for Europeans (Section 3.3). We repeated inference of the same parameters using the same ESP data. The inferred parameters for *exponential* growth are also similar to those obtained in [127] that were based on the model by [49] (Table 3.2). However, the SFS from this model fits the data worse than that from the *exponential* model based on the ancient history of the model by [44] (p-value $= 1.59 \times 10^{-6}$ from χ^2 goodness of fit test between *exponential* Gravel model and ESP data, which is 0.97 for *exponential* Gazave model; see Appendix A.12 and Table 3.3). By applying a *generalized* growth epoch to the model by [49], the inferred parameters are generally in line with those from *generalized* Gazave model, although some differences exist (Table 3.2), indicating that the assumption of ancient history can affect the inference of recent growth to some extent. More importantly, *generalized* Gravel model fits the data almost equally well as *generalized* Gazave model, which is significantly better than the *exponential* model (p-value $\ll 10^{-12}$ by

χ^2 likelihood-ratio test; also see Table 3.3). As with *generalized* Gazave model, the inferred growth speed parameter from *generalized* Gravel model, $b = 1.22$ (1.18-1.26), is also significantly larger than the exponential speed $b = 1$ (P-value $\ll 10^{-12}$ using one-tailed z-test; Figure 3.7(CD)).

Table 3.3: **Goodness of fit between the SFS from inferred models and ESP data.** We show the p-value from χ^2 goodness of fit test and KL divergence between the SFS from the ESP data and that from the constant population size model, the inferred exponential model, the generalized model and the two-epoch exponential model. The assumed ancient history (Gazave model or Gravel model) is indicated in parenthesis. The constant population size model is included here for comparison purposes.

Model	p-value from χ^2 test	KL divergence
Constant	0	0.84
Exponential (Gazave)	0.97	1.64×10^{-4}
Generalized (Gazave)	1	1.15×10^{-4}
Two-Epoch Exponential (Gazave)	1	1.09×10^{-4}
Exponential (Gravel)	1.59×10^{-6}	4.12×10^{-4}
Generalized (Gravel)	1	1.15×10^{-4}

Motivated by these results, we considered a third model with *two* recent exponential growth epochs, which still assumes the ancient epochs before 620 generations ago to be the same as those in the model by [44] illustrated in Figure 3.2(A). Five parameters were inferred (Table 3.4), with the first phase of growth estimated to start 219 (95-334) generations ago with a population size of 12,200 (11,700-13,200). This phase of growth lasts until 135 (25-157) generations ago and leads to a population size of 47,100 (30,200-540,900). The population size after the recent phase of growth is 1.12 (1.07-2.09) million. This model provides a significantly better fit than the model with a single *exponential* growth (P-value $= 5.55 \times 10^{-6}$ by χ^2 likelihood-ratio test with 2 degrees of freedom), but is a worse model than the *generalized* growth model (based on

Bayesian Information Criterion, $\mathrm{BIC}_{\mathrm{two-epoch-exponential}} - \mathrm{BIC}_{\mathrm{generalized}} = 6.1$). However, this model exhibits some of the same accelerating pattern as in the *generalized* growth model, ascertained by the growth rate of the most recent exponential epoch being 2.4% (2.3%-5.2%), larger than that of the first exponential epoch, 1.6% (1.3%-2.1%). This acceleration pattern shown in both the generalized model and the model with two exponential epochs is consistent with evidence of growth in European census population size that has greatly accelerated in Modern Era [67].

Table 3.4: **Demographic inference results using ESP data for a model with two recent epochs of exponential growth.** Shown are point estimates and 95% confident intervals (in parenthesis) for the following parameters of the inferred epoch: population size before growth (N_2), population size after the more ancient phase of exponential growth (N_1), population size after the recent phase of exponential growth (N_0), time when the ancient phase of exponential growth started (T_2, in generations), time when the recent phase of exponential growth started (T_1, in generations).

$N_2(10^4)$	$N_1(10^4)$	$N_0(10^6)$	T_2	T_1
1.22	4.71	1.12	219	135
$(1.17 \sim 1.32)$	$(3.03 \sim 54.09)$	$(1.07 \sim 2.09)$	$(95 \sim 334)$	$(25 \sim 157)$

3.5 Discussion

In this study, we provide the mathematical derivation and a software that can efficiently compute the expected values of five genetic data summary statistics given a generalized demographic model by evaluating the derived explicit expressions. These summary statistics include the time to the most recent common ancestor (T_{MRCA}), the total number of segregating sites (S), the site frequency spectrum (SFS), the average pairwise difference between chromosomes per site

(π) and the burden of private mutation (α). The fast and accurate generation of these summary statistics under generalized models can provide a useful tool in the studies of human demographic inference. For instance, in addition to inference based on the SFS as in the present study, a recent study [22] presented an inference framework based on the total number of segregating sites. The results in this study can be easily incorporated into that framework. Furthermore, the source code of the software is freely available to allow extensions to compute other summary statistics of interest (for example, the joint-SFS of samples from multiple populations under generalized models by extending [21] and [137]). Such extensions can facilitate a variety of population genetic studies in humans and other organisms beyond the inference of demographic history.

It is also possible that other families of growth models may fit the pattern of human population size history. For instance, a recent study [34] considered the algebraic-growth model in the form of $N(T) = T^{\gamma}$. In reality, however, not all demographic models have numerically stable closed-form expressions for the expected time of the first coalescent event (ψ_j). In these cases, fast and accurate numerical integration methods, such as Gauss-Legendre quadrature used in this work, can be applied to evaluate ψ_j. This technique holds the promise of efficiently generating the expected value of population genetic summary statistics under arbitrary population size functions.

It has been pointed out that as sample size increases, the assumptions of standard Kingman's coalescent are violated as multi-merger and simultaneous-merger events can become non-negligible [6]. Such events can distort the genealogies and potentially cause the values of summary statistics to be different from those under Kingman's coalescent [6]. To explore such discrepancies, we

compared the SFS from Kingman's coalescent and discrete-time Wright-Fisher (DTWF) model [6] under the inferred demographic history in *generalized* Gazave model with a sample size of 3,870 diploids (Appendix A.13). We observed that the SFS from DTWF model and Kingman's coalescent are very similar (Figure 3.8), which means that multi-merger and simultaneous-merger events should not have a significant effect on the inference carried out in this study. However, it remains valuable to systematically study the effect of multi-merger and simultaneous-merger events in the context of generalize growth, especially as sample size increases.

By applying inference of generalized growth on the SFS generated from the synonymous variants of 4,300 individuals of NHLBI ESP dataset [41, 127], we found that *generalized* growth model shows a better fit to the observed data than the *exponential* growth model that has been used by all previous demographic modeling studies (P-value $= 3.85 \times 10^{-6}$). We also find that the European population experiences a recent growth in population size with speed modestly faster than exponential ($b = 1.12$, P-value $\ll 10^{-12}$ for difference from $b = 1$). This result is consistent with previous speculations that human population might have undergone a recent accelerated growth epoch based on the observation of very rare, previously unknown variants in several sequencing studies with large sample sizes [41, 108, 127]. It is also in line with the super-exponential growth in census population size during that time [67]. In future studies, it will be valuable to incorporate gradient-based optimization techniques for the fast inference of demographic models containing generalized growth epochs, e.g., by extending the work of [8]. Such improvement will enable simultaneous inference of recent growth and more ancient epochs.

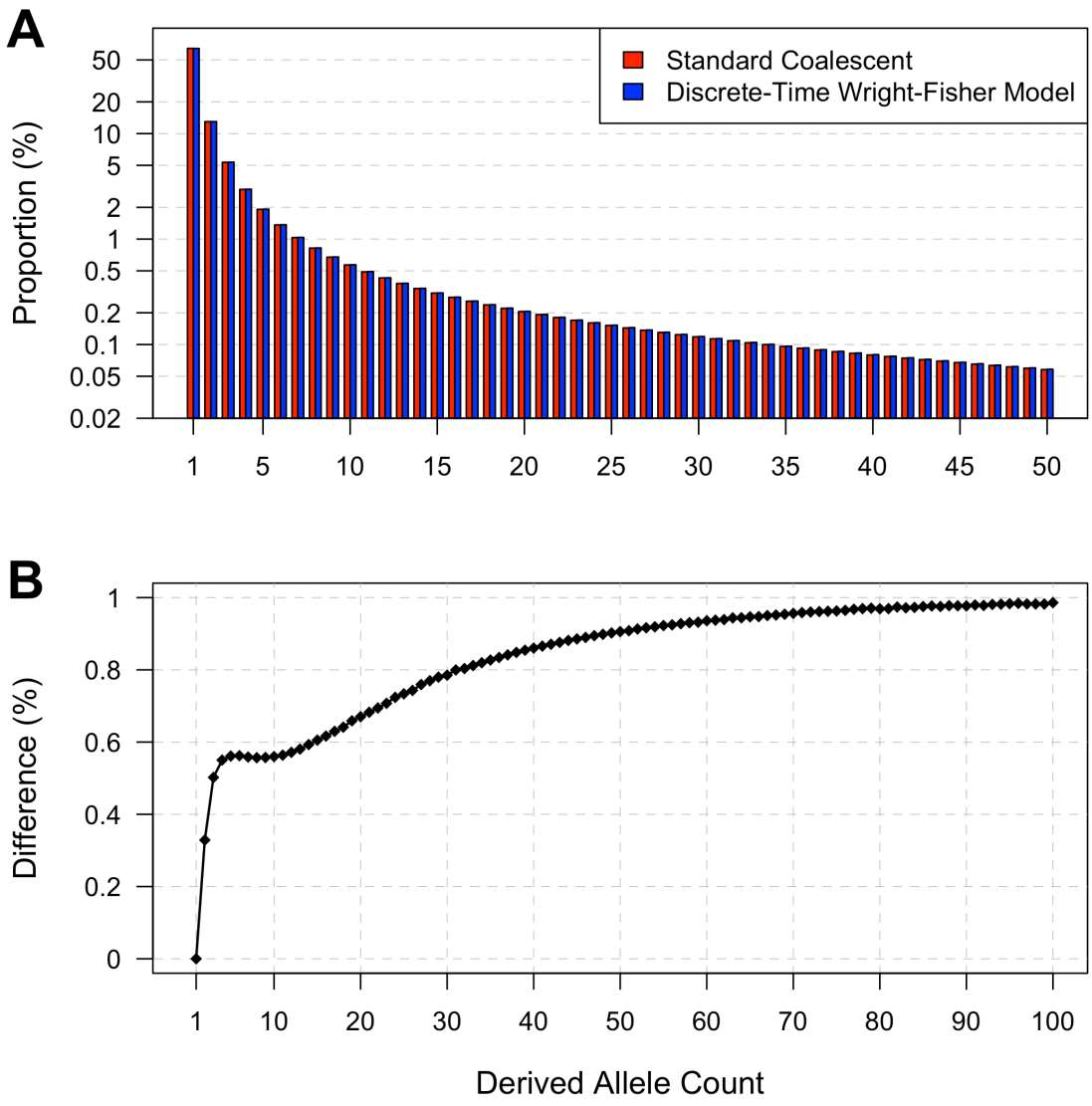

Figure 3.8: **Effects of multi-merger and simultaneous-merger events on the SFS.** The underlying demographic model is the best-fit generalized model using the ancient history in Gazave *et al*. (2014). The sample size is 3,870 diploid individuals. (A) The 100-entry *partially normalized* SFS under Kingman's coalescent and under discrete-time Wright-Fisher model. (B) The percentage difference of entry-to-singleton ratio between Kingman's coalescent and discrete-time Wright-Fisher model for the first 100 entries.

To minimize the impact of natural selection on our demographic inference, we considered only synonymous SNVs for demographic modeling, as in the original study [127]. However, it is still a potential limitation that the data is affected by negative and background selection. Hence, it remains valuable to validate the result of super-exponential growth by conducting inference on SFS calculated from more neutral genomic regions [44] or by modeling the effect of selection. One promising possibility is extracting genomic regions that are less subject to selection from whole genome sequences in the UK10K project [126]. More generally, with the increasing availability of high-quality whole-genome sequencing data with large sample sizes for humans and other species, more refined and realistic demographic models can be estimated with generalized growth models.

CHAPTER 4

NEW METHODS FOR ANALYZING THE X-CHROMOSOME IN ASSOCIATION STUDIES WITH SOFTWARE IMPLEMENTATION AND APPLICATIONS

4.1 Abstract

XWAS is a new software suite for the analysis of the X chromosome in association studies and similar genetic studies. The X chromosome plays an important role in human disease and traits of many species, especially those with sexually dimorphic characteristics. Special attention needs to be given to its analysis due to the unique inheritance pattern, which leads to analytical complications that have resulted in the majority of genome-wide association studies (GWAS) either not considering X or mishandling it with toolsets that had been designed for non-sex chromosomes. We hence developed XWAS to fill the need for tools that are specially designed for analysis of X. Following extensive, stringent, and X-specific quality control, XWAS offers an array of statistical tests of association, including: 1) the standard test between a SNP (single nucleotide polymorphism) and disease risk, including after first stratifying individuals by sex, 2) a test for a differential effect of a SNP on disease between males and females, 3) motivated by X-inactivation, a test for higher variance of a trait in heterozygous females as compared with homozygous females, and 4) for all tests, a version that allows for combining evidence from all SNPs across a gene. We applied the toolset analysis pipeline to 16 GWAS datasets of immune-related disorders and 7 risk factors of coronary artery disease, and discovered several new X-linked genetic associations. XWAS will provide the tools and incentive for others to

incorporate the X chromosome into GWAS and similar studies in any species with an XX/XY system, hence enabling discoveries of novel loci implicated in many diseases and in their sexual dimorphism.

4.2 Introduction

Genome-wide association studies (GWAS) have identified thousands of loci underlying complex human diseases and other complex traits [138]. While successful for the autosomes (nonsex chromosomes), the vast majority of these studies have either incorrectly analyzed or ignored the X chromosome (X) [142]. In most studies, all variants on the X have been removed as a consequence of the quality control (QC) procedures [18, 97, 132, 142]. Many other studies that did analyze the X chromosome incorrectly applied methods that have been designed for the autosomes, without accounting for the analytical problems arising from X's unique mode of inheritance and its consequent population genetic and evolutionary patterns [4, 35, 51, 52, 70, 80, 89, 141]. As a result, the role X plays in complex diseases and traits remains largely unknown.

Many human diseases commonly studied in GWAS show sexual dimorphism, including autoimmune diseases [136], cardiovascular diseases [81] and cancer [100, 106], which suggests a potential contribution of the X chromosome [17, 109]. Several recent studies have examined this issue and demonstrated the potential value of analyzing X [18, 46, 84, 94, 133]. However, while association methods, QC and analysis pipelines are well established for the autosomes, respective pipelines for X-linked data are not readily available. Hence, in this article, we introduce the software package XWAS (chromosome X-Wide Anal-

ysis toolSet), which is tailored for analysis of genetic variation on X in human and other species with an XX/XY system. It implements extensive functionality that carries out QC specially designed for the X chromosome, statistical tests of single-marker association that account for its unique mode of inheritance, gene-based tests of association, and additional distinct tests only applicable to X that capitalize on its mode of inheritance. In implementing these features, the toolset builds on–and complements–the commonly used PLINK [113] software. It includes many novel features that can facilitate X-wide association studies that are not available in PLINK and, to the best of our knowledge, in any other software. Combined, the XWAS toolset integrates the X chromosome into GWAS as well as into the next generation of sequence-based association studies and into studies of other species.

4.3 Features and Functionality

4.3.1 Quality Control Procedures

The XWAS toolset implements a whole pipeline for performing QC on genotype data for the X chromosome. The pipeline first follows standard GWAS QC steps as implemented in PLINK [113] and SMARTPCA [112] by running these tools. These include the removal of both individual samples and SNPs (single nucleotide polymorphisms) according to multiple criteria. Specifically, samples are removed based on 1) relatedness, 2) high genotype missingness rate, and 3) genetic ancestry differing from the majority of the samples [112]. SNPs are removed based on criteria such as their missingness rate, their minor allele fre-

quency (MAF), and deviation from Hardy-Weinberg equilibrium (HWE). This entire QC pipeline is applicable to both case-control GWAS (binary traits) and GWAS of quantitative traits. One filter applied only to binary traits is the removal of SNPs for which missingness is correlated with the trait, that is, with case or control status (*--test-missing*).

To consider differences in genotyping between hemizygous males and diploid females, XWAS applies all the aforementioned QC steps of samples separately for males and females. Consequently, a unified dataset is generated for subsequent analyses that include all SNPs and individuals passing the above filtering criteria in both the male and female QC groups.

The pipeline then applies X-specific QC steps, which are exclusively built into XWAS, to the unified dataset. These include 1) removing SNPs with significantly different MAF between male and female samples in the control group (*--freqdiff-x*), 2) removing SNPs with significantly different missingness rates between male and female controls (*--missdiff-x*), and 3) the removal of SNPs in the pseudoautosomal regions (PARs). The first 2 of these steps capture problems in genotype calling when plates include both males and females [76].

4.3.2 Single-Marker Association Testing on the X Chromosome

For an X-linked SNP, while females have 0, 1, or 2 copies of an allele, hemizygous males have at most one copy. Via the process of X-inactivation, 1 of the 2 copies in females is usually transcriptionally silenced. If X-inactivation is complete, it produces monoallelic expression of X-linked protein-coding genes in females. Therefore, when considering loci that undergo complete X-inactivation,

it may be apt to consider males as having $0/2$ alleles, corresponding to the female homozygotes (the FM_{02} test). The toolset carries out this test for association between a SNP and disease risk by using the *--xchr-model 2* option in PLINK [113]. For other scenarios though, including where some genes on the X escape X-inactivation or different genes are inactivated in different cells, it can be more indicative to code males as having $0/1$ alleles. Hence, the toolset further carries out such an association test (FM_{01} test) of a SNP by using the following options in PLINK : *--logistic* and *--linear* for binary and quantitative traits, respectively.

All tests, including tests described in following sections, allow for covariates such as population structure, sex, and traits that are correlated with the disease, as commonly considered in GWAS. We suggest calculating principal components by using EIGENSTRAT [112] and include them as covariates to control for population structure. Ten such principal components are considered by default, unless otherwise specified. Any other user-defined covariates can also be incorporated.

4.3.3 Single-Marker Sex-Stratified Analysis on the X Chromosome

The XWAS software further includes new tests that are not included in PLINK. First, we implemented a new sex-stratified test, FM_{comb}, which is particularly relevant for X analyses since SNPs and loci on the sex chromosomes are potentially more likely to exhibit different effects on disease risk between males and females. In such scenarios, as well as in scenarios where the effect is only observed in one sex, a sex-stratified test as described in the following can be

better powered. This functionality is accessible by the option *--strat-sex*. The FM_{comb} test first carries out an association test separately in males and females and then combines the results of the 2 tests to obtain a final sex-stratified significance level. The combination of the 2 test statistics is implemented using both Fisher's method (*--fishers*) [39] (in the $FM_{F.comb}$ test) and Stouffer's method (*--stouffers*) [123] (in the $FM_{S.comb}$ test).

Each of these 2 tests is more powerful in different scenarios [18], for example, $FM_{F.comb}$ allows the SNP tested to have different, even an opposite, effect on disease risk in males and females. $FM_{F.comb}$ is also insensitive to whether males are coded as $0/2$ (as in the FM_{02} test) or as $0/1$ (as in the $FM_{S.comb}$ test), thus making no assumptions regarding X-inactivation status. Alternatively, $FM_{S.comb}$ directly accounts for the potentially differing sample sizes between males and females to maximize power. For this latter test, XWAS weighs by the sample size in males and females in cases and controls following the approach of [139].

4.3.4 Single-Marker Sex-Differentiated Effect Size Test on the X Chromosome

We described above sex-stratified tests that accommodate associations with different effect size between males in females. In another type of test (FM_{diff}), we directly test whether the effect size is different between the sexes. This test, applied to each SNP, runs a *t*-test [116] to test for difference between the odds ratio (OR) in males alone and the OR in females alone for case-control studies,

$$t = \frac{\log \mathrm{OR}_{\mathrm{male}} - \log \mathrm{OR}_{\mathrm{female}}}{\sqrt{se^2_{\mathrm{male}} + se^2_{\mathrm{female}} - 2r \cdot se_{\mathrm{male}} \cdot se_{\mathrm{female}}}} \tag{4.1}$$

or to test for difference the male-specific and female-specific β estimates for quantitative traits,

$$t = \frac{\beta_{\text{male}} - \beta_{\text{female}}}{\sqrt{se^2_{\text{male}} + se^2_{\text{female}} - 2r \cdot se_{\text{male}} \cdot se_{\text{female}}}} \tag{4.2}$$

while accounting for hemizygosity in males. This test is implemented under the *--sex-diff* option. For this test and the sex-stratified test, both odds ratios and regression coefficients in each sex can be provided as output for further examination.

4.3.5 Single-Marker Variance-Based Testing Informed by X-Inactivation in Females

The details of this variance-based test are described in [84]. The major work of this test was done by the authors of that paper.

During X-inactivation, the expression of one copy of the X chromosome in females is randomly silenced, thereby increasing variation in the expression of X-linked quantitative trait loci (QTL). Specifically, females that are heterozygotes for a QTL might exhibit higher phenotypic variance than homozygous females since one or the other allele might be more dominantly affecting the phenotype in each given female heterozygote, such that for some individuals the QTL expression is more similar to one type of female homozygous, while to the other type in other individuals. We developed a test aimed at capturing this increased variance as a means for detecting X-linked QTLs in females. This test (F_{var}) is implemented under the *--var-het* option. Although this F_{var} test is implemented for quantitative traits, it can be generalized to qualitative traits by applying liability threshold modeling [144] to transform disease status to an unobserved

continuous liability.

The null hypothesis of the F_{var} test is that phenotypic variances of the 3 genotypic groups of a SNP with 0, 1, or 2 copies of an allele are all equal. The alternative hypothesis is that female heterozygotes show a higher phenotypic variance than others. Hypothesis testing is carried out using a modified Brown-Forsythe test of variances [15]. We first normalize the phenotypic value and remove the effects of possible covariates by a linear regression as conventionally done, namely

$$y = \mu + XB + \epsilon, \tag{4.3}$$

where y is a vector of quantitative trait levels, μ is the population mean, X is the matrix of possible covariates, and ϵ is a vector of residuals. Assume $y_{i|g=j}$ is the phenotypic value of the i^{th} individual in the j^{th} genotypic group and $z_{i|g=j} = |\epsilon_{i|g=j}|$ is the absolute residual value of the i^{th} individual in the j^{th} genotypic group ($j = 0, 1$, or 2 copies of an allele of a SNP). A test statistic is derived as

$$T_{\text{var}} = \frac{\bar{z}_1 - \bar{z}_{0/2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_{0/2}^2}{n_0+n_2}}} \tag{4.4}$$

where $\bar{z}_1$ is the sample mean of $z_{i|g=1}$ over i, $\bar{z}_{0/2}$ is the sample mean of $z_{i|g=0}$ and $z_{i|g=2}$ combined, s_1^2 and $s_{0/2}^2$ are the sample variances, respectively, and n_j is the sample size of the j^{th} genotypic group. Under the null hypothesis, the statistic follows a t-distribution with degrees of freedom given by

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_{0/2}^2}{n_0+n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1} + \frac{\left(s_{0/2}^2/(n_0+n_2)\right)^2}{n_0+n_2-1}} \tag{4.5}$$

This variance-based test captures a novel signal of X-linked associations by directly testing for higher phenotypic variance in heterozygous females than homozygotes. As a test of variance it is generally less powerful than standard

tests of association that consider means; however, it provides an independent and complementary test to the standard association test for QTLs on X [94].

4.3.6 X-Linked Gene-Based Testing

XWAS also includes unique features for carrying out gene-based association analysis on the X chromosome. Gene-based approaches may be better powered to discover associations than single-marker analyses in cases of a gene with multiple causal variants of small effect size, or of multiple markers that are each in incomplete linkage disequilibrium with underlying causal variant/s. Furthermore, in studying the effect of X on sexual dimorphism in complex disease susceptibility, it may be desirable to analyze whole-genes or all genes of a certain function combined based on their unique function or putatively differential effect between males and females [18].

The XWAS package determines the significance of association between each gene as a whole and disease risk by implementing a gene-level test statistic that combines individual SNP-level test statistics for all SNPs in and around each studied gene. This gene-level approach is applicable to any of the different tests described above. For instance, beyond tests of association, it can be applied to the sex-differentiated tests. In this case the gene-based test captures any scenario whereby SNPs within the gene display different effects in males and females, without restricting such differential effects to be of a similar nature across SNPs. By default, genes are considered from the UCSC browser "knownCanonical" transcript ID track. SNPs are mapped to a gene if they are in the gene or within 15kb of the gene's start or end positions. The user can also provide a

different set of gene definitions or alternate regions of interest and a different window length around them in which SNPs are also to be considered.

Combining SNP statistics across a gene is implemented in the general framework of [86]. Specifically, the significance of an observed gene-based test statistic is assessed from the distribution of test statistics that is expected given the linkage disequilibrium between the SNPs in the gene. In [86], the test statistics for all SNPs in the gene are summed. Here, we have implemented a slight modification to this procedure, whereby we combined SNP-based P-values with either the truncated tail strength [58] or the truncated product [145] method, which have been suggested to be more powerful in some scenarios [93, 145].

To determine significance, XWAS follows the procedure in [86]. The observed statistic is compared to gene-level test statistics obtained when SNP-level statistics are randomly drawn from a multivariate normal distribution with a covariance determined by the empirical linkage disequilibrium between the SNPs in the tested gene. The significance level is then the proportion, out of many such drawings, for which this sampled gene-level statistic is more, or as extreme compared with the empirical one. For computational efficiency, the number of drawings is determined adaptively as in [86]. By combining the truncated tail measures with this procedure, our new gene-based method combines the test statistics from multiple SNPs that show relatively low *P*-values, while also accounting for the dependency between these *P*-values due to linkage disequilibrium between the SNPs. Such a gene-based *P*-value is estimated for each gene and for each of the X-linked tests described above.

4.4 Implementation

The XWAS software package is implemented in C++ and includes in part functions from open-source PLINK [113]. XWAS uses the same input format as PLINK. Beyond C++, additional features are also implemented in scripts, including in shell (for QC), Perl (for converting file formats and using SMARTPCA), and R (for gene-based testing). The entire package is freely available for download from http://keinanlab.cb.bscb.cornell.edu and includes 1) scripts, 2) the binary executable XWAS, 3) all source code with a Makefile, 4) a user manual, and 5) example data and examples of running the different options offered by the package. Additional help is provided via the *--xhelp* option. The XWAS toolset was initially designed and optimized for Linux systems, hence exhibits best performance in such systems. A Makefile is also provided to facilitate local compilation on Linux environments, and can also be adjusted for Windows and MAC OS by revising a few lines indicated therein.

4.5 Applications

In this section, we summarize several sets of results obtained by applying the XWAS software to several publicly available GWAS datasets. These results are intended to demonstrate the usefulness of the XWAS software. More detailed results are described in other papers [18, 94].

4.5.1 Association of X-Linked SNPs with Autoimmune Diseases

We applied the XWAS software to 16 GWAS datasets of autoimmune disease and other disorders with a potential autoimmune-related component. These include the following datasets that we obtained from dbGaP (Mailman et al. 2007; Tryka et al. 2014): ALS Finland [78] (phs000344), ALS Irish [30] (phs000127), Celiac disease CIDR [3] (phs000274), MS Case Control (Baranzini et al. 2009) (phs000171), Vitiligo GWAS1 [61] (phs000224), CD NIDDK [33] (phs000130), CASP [107] (phs000019), and T2D GENEVA [114] (phs000091). Similarly, we obtained the following datasets from the Wellcome Trust Case Control Consortium (WT): all WT1 (The Wellcome Trust Case Control Consortium 2007) datasets, WT2 ankylosing spondylitis [36], WT2 ulcerative colitis [134] and WT2 multiple sclerosis [57]. Finally, we also analyzed data from Vitiligo GWAS2 [60].

Following application of the QC pipeline as described above, we applied the SNP-level FM_{02}, $FM_{F.comb}$, and $FM_{S.comb}$ tests to all SNPs in each of the 16 datasets. Based on the Vitiligo GWAS1 datasets, we associated SNPs in a region 17 kilobases (kb) away from the retrotransposed gene retro-*HSPA8* with risk of vitiligo. The parent of this retrotransposed gene, *HSPA8* on chromosome 11, encodes a member of the heat shock protein family, which has been previously associated to vitiligo [1, 104, 105]. We discovered another association in WT2 ulcerative colitis of SNPs in an intron of *BCOR* contributing to ulcerative colitis disease risk. *BCOR* indirectly mediates apoptosis via co-repression of *BCL-6* [56].

4.5.2 Association of Whole X-Linked Genes with Autoimmune Diseases

We next focused on a gene-based analysis of the X chromosome by using the SNP-level results of all the 3 tests in the above results as a basis for gene-based tests in the same 16 datasets. This analysis led to the discovery of the first X-linked gene-based associations with any disease or trait, which supports the utility of the XWAS package in facilitating such analyses. We associated in Vitiligo GWAS1 and replicated in Vitiligo GWAS2 an association between the gene *FOXP3* and vitiligo disease risk, in support of an earlier candidate gene study [10]. We also found a novel association of *ARHGEF6* to Crohn's disease and further replicated it in ulcerative colitis, another inflammatory bowel disorder (IBD). *ARHGEF6* binds to a surface protein of a gastric bacterium (Helicobacter pylori) that has been associated to IBD [59, 91]. Finally, we associated *CENPI* as contributing to the risk of 3 different diseases (amyotrophic lateral sclerosis, celiac disease, and vitiligo). Other, autosomal genes in the same family as *CENPI* have previously been associated to amyotrophic lateral sclerosis [2] as well as multiple sclerosis [5], supporting an involvement of *CENPI* with autoimmunity in general.

4.5.3 X-Linked SNPs Showing Sex-Differentiated Effect Size with Autoimmune Disease

As a final analysis on the 16 autoimmune datasets, we applied the FMdiff test and its gene-based version. Based on this test, we discovered and replicated the

gene *C1GALT1C1* (also known as *Cosmc*) as exhibiting sex-differentiated effect size in risk of IBD. *C1GALT1C1* is necessary for the synthesis of many O-glycan proteins [65], which are components of antigens. We further found *CENPI*, which we previously associated with several diseases, to show significantly different effects in males and females in the same diseases as in the association analysis.

4.5.4 Increased Variance of Systolic Blood Pressure in Heterozygous Females for an X-Linked SNP

These results are described in [84] in detail. The major work was done by the authors of that paper.

As an example application of the variance-based testing (F_{var}) informed by X-inactivation, we considered data on 7 quantitative traits from the Atherosclerosis Risk in Communities (ARIC) study [140] along with Affymetrix 6.0 data from the participating individuals, which included 34,527 X-linked SNPs. First, we applied the entire set of QC procedures implemented in XWAS for quantitative traits. Then, we applied our single-marker variance-based testing and compared with application of standard testing for a QTL. Across the 7 traits, we found 1 SNP with a significant association based on the variance test [94]. Importantly, the signals of this test are not in the same loci as those of the standard test, in line with them capturing different types of signals. Specifically, the significant SNP, rs4427330, which is associated with systolic blood pressure based on the variance test, is not associated with any trait based on the standard test. It is located upstream of *AFF2*, which might regulate *ATRX*. *ATRX* is as-

sociated with alpha-thalassemia, a disease that can cause anemia and has been associated with hypertension [11].

4.6 Conclusions

We have developed XWAS, an extensive toolset that facilitates the inclusion of the X chromosome in GWAS. It offers X-specific QC procedures, a variety of X-adapted tests of association, and an X-specific test of variance testing, available for both single-marker and gene-based statistics. We applied this toolset to successfully discover and replicate a number of genes with autoimmune disease risk and blood pressure.

Additional changes and optimizations will be included in future versions of this software. While imputation of unobserved SNPs is presently performed as a preprocessing step using IMPUTE2 [54], we will incorporate X-specific imputation as part of the pipeline. Additional features will include analysis of X-linked data from sequence-based association studies (including burden tests), statistical methods that have been previously designed for the X chromosome [24, 25, 90, 130, 147], additional tests we previously proposed based on the workings of X-inactivation [94], and tests for gene–gene interactions. Finally, we will incorporate information regarding whether or how often a gene undergoes or escapes X-inactivation [17, 28, 32]. For computational efficiency, we will also continually upgrade the functions of PLINK that XWAS uses to the most recent version.

This software, alone and through incorporation of additional features, can be used for other types of studies of the X chromosome beyond association studies,

in particular population genetic studies. For instance, allele frequency output and testing for significant differences in allele frequency between males and females as currently implemented, can be utilized to search for signals of selection.

Considering the availability of unutilized data for the X chromosome from thousands of GWAS, and the additional X-linked data that is being generated as part of ongoing GWAS, many researchers will find extensive utility in the XWAS toolset. Furthermore, it is not limited to application to human data, but rather genetic data from all organisms with XX/XY sex determination system, including all mammals. XWAS will facilitate the proper analysis of these data, incorporate X into GWAS and enable discoveries of novel X-linked loci as implicated in many diseases and in their sexual dimorphism.

APPENDIX A

APPENDIX FOR CHAPTER 3

A.1 Detailed description of genetic summary statistics

A.1.1 Total number of segregating sites (S)

Suppose we have n sequences (chromosomes), this quantity stands for the number of sites in which the sequences have different genotypes. Namely, if all sequences have a common genotype for a site, this site is not considered as a segregating site.

A.1.2 Time to the most recent common ancestor (T_{MRCA})

This statistic is the time taken for all of the samples at present to coalesce to the same ancestor.

A.1.3 Site frequency spectrum (SFS)

Suppose we have n sequences sampled at present, the full SFS $\boldsymbol{\xi}$ has $(n-1)$ entries $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_{n-1})$, where ξ_i records the fraction of segregating sites that have i derived alleles and $(n-i)$ ancestral alleles. When we don't have information about the ancestral allele, the folded SFS $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_{\lfloor \frac{n}{2} \rfloor})$ is used, where η_i records the fraction of segregating sites that have i minor alleles and $(n-i)$ major alleles. By definition, $\eta_i = \frac{\xi_i + \xi_{n-i}}{1+\delta(i,n-i)}$.

A.1.4 Average pairwise difference per site (π)

Suppose we have n sequences sampled at present. We compare every two different sequences (thus there are $\binom{n}{2}$ pairs), count the number of differences between each pair, calculate the average of the total differences and normalize the average difference by the total number of sites, or total length of loci L. This quantity has the following relationship with the SFS and S:

$$\pi = \frac{S}{L\binom{n}{2}} \sum_{i=1}^{n-1} i(n-i)\xi_i = \frac{S}{L\binom{n}{2}} \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} i(n-i)\eta_i. \tag{A.1}$$

A.1.5 Burden of private mutations (α)

Suppose we have n diploid individuals sequenced (thus there are $2n$ sequences). α stands for the proportion of heterozygous positions in a newly sequenced $(n+1)^{\text{th}}$ individual that are novel. Namely, all of the previous n individuals have the same genotype at such a site, but this newly sequenced individual have a different genotype.

A.2 More detailed explanation of the growth speed parameter b_k

When $r_k \neq 0$, the growth speed is controlled by the parameter b_k. With the same value of r_k, $N_{k,\text{f}}$ and $(T_k - T_{k-1})$, if $b_k > 1$, the model will reach a $N_{k,\text{i}}$ larger than that of an exponential model. As a result, it is considered to be faster than exponential or super-exponential. Similarly, if $b_k < 1$, the model will reach a $N_{k,\text{i}}$

smaller than that of an exponential model and thus is considered to be slower than exponential or sub-exponential.

To illustrate the above facts, we give an example in Figure A.1(A). The growth epoch starts 200 generations ago with a population size of 10,000. The value of growth rate $g = \frac{d}{dT} \log N(T)$ is fixed at 0.35% such that when exponential growth model is used, the population size after growth is 20,000, which is a 2-fold growth. The values of b are chosen to be 0.9, 1 and 1.1. When $b = 1.1$, the population size after growth is 67,730, larger than 20,000 when exponential growth is considered. Similarly, when $b = 0.9$, the population size after growth 13,129, smaller than 20,000.

If we fix $N_{k,\mathrm{i}}$, $N_{k,\mathrm{f}}$ and $(T_k - T_{k-1})$, as is mostly considered in this study, taking different values of b will cause the growth pattern to be different. When $b > 1$, the growth will show an accelerating pattern compared with exponential growth; while when $b < 1$, the growth will show a decelerating pattern. To illustrate this point, consider the models shown in Figure A.1(B). The growth epoch is from 200 generations ago to present and the population sizes before and after growth are fixed at 10,000 and 100,000 respectively. The values of b are chosen to be 0.3, 1 and 1.7. For the exponential model, the growth rate 1.15% is constant throughout the epoch. For $b = 1.7$, the growth rate (0.52%) is smaller than that of the exponential growth (1.15%) at the onset time of 200 generations ago. The growth keeps accelerating as time approaches present. At $t = 0$, the growth rate for $b = 1.7$ (2.87%) is larger than that of the exponential (1.15%). For $b = 0.3$, the pattern is opposite. The instantaneous growth rate (2.87%) is larger than that of the exponential growth (115.13) at 200 generations ago. The growth keeps decelerating as time approaches present. At $t = 0$, the instantaneous

growth rate for $b = 0.3$ (0.57%) is smaller than that of the exponential (1.15%).

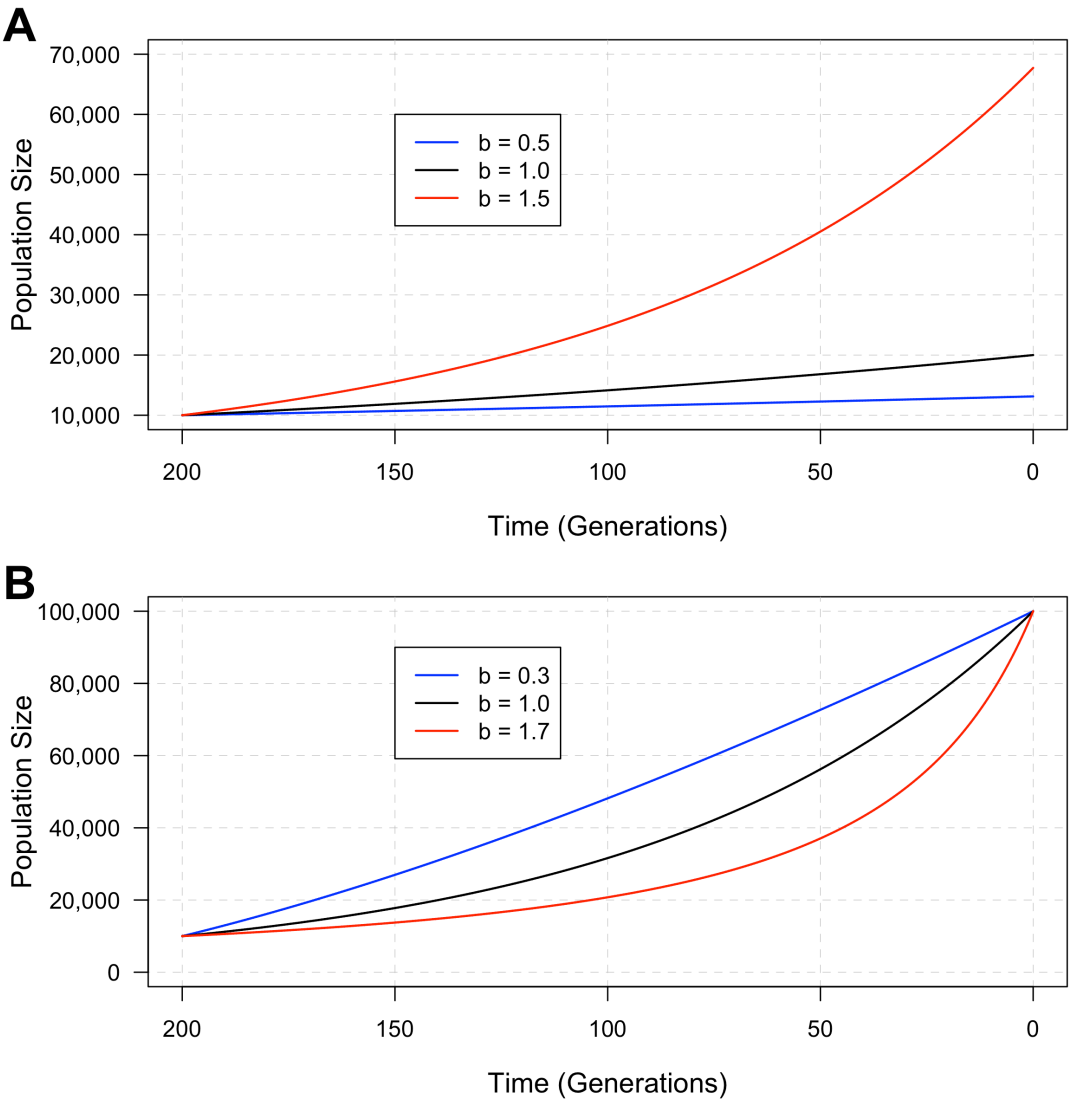

Figure A.1: **Different patterns of generalized growth.** (A) Illustration of the population size functions when keeping the population size before growth N_f, the growth time T and the parameter r the same and varying the growth speed parameter b to be 0.5, 1.0 and 1.5. (B) Illustration of the population size functions when keeping the population size before growth N_f, the population size after growth N_i and the growth time T the same and varying the growth speed parameter b to be 0.3, 1.0 and 1.7.

A.3 Quantities A_j^p, V_j^p and $W_{i,j}^p$

For computing $\mathbb{E}[T_{\text{MRCA}}^p]$, the quantities A_j^p can be calculated by [111, 124, 125]

$$A_j^p = \frac{(-1)^j(2j-1)p_{[j]}}{p^{(j)}}, \tag{A.2}$$

where $p_{[j]}$ is the falling factorial function, $p_{[j]} = p(p-1)\cdots(p-j+1)$, and $p^{(j)}$ is the rising factorial function, $p^{(j)} = p(p+1)\cdots(p+j-1)$.

For computing $\mathbb{E}[\boldsymbol{\xi}^p]$, the quantities V_j^p can be calculated by [110]

$$V_j^p = (2j-1)\frac{p!(p-1)!}{(p+j-1)!(p-j)!}[1+(-1)^j], \tag{A.3}$$

and $W_{i,j}^p$ are constants given by the following recursive relationships [110]:

$$W_{i,2}^p = \frac{6}{p+1}; \tag{A.4}$$

$$W_{i,3}^p = \frac{30(p-2i)}{(p+1)(p+2)}; \tag{A.5}$$

$$W_{i,j+2}^p = -\frac{(1+j)(3+2j)(p-j)}{j(2j-1)(p+j+1)}W_{i,j}^p + \frac{(3+2j)(p-2i)}{j(p+j+1)}W_{i,j+1}^p. \tag{A.6}$$

A.4 Expressions of r_k

For the generalized growth models considered in this study, any epoch k is determined by the starting population size $N_{k,\text{i}}$, the ending population size $N_{k,\text{f}}$, the duration of the epoch $(T_k - T_{k-1})$ and the growth speed parameter b_k. After determining the epoch, the dependent parameter $r_k = r_k(N_{k,\text{i}}, N_{k,\text{f}}, b_k, T_k - T_{k-1})$ is calculated by

$$r_k = \begin{cases} \frac{N_{k,\text{i}}^{1-b_k} - N_{k,\text{f}}^{1-b_k}}{T_k - T_{k-1}}, & b_k \neq 1 \\ \frac{\log N_{k,\text{i}} - \log N_{k,\text{f}}}{T_k - T_{k-1}}, & b_k = 1 \end{cases}. \tag{A.7}$$

A.5 Expressions of $\Lambda(T)$ for evaluating ϕ_j^k

For convenient purposes, define $\lambda_k(T) = \int_{T_{k-1}}^{T} d\sigma/\mathcal{N}(\sigma)$, where $\mathcal{N}(\sigma) = 2N(\sigma)$ and $T_{k-1} \leq T \leq T_k$, then $\Lambda(T) = \Lambda(T_{k-1}) + \lambda_k(T)$. For generalized models, the solution for $\lambda_k(T)$ is

$$\lambda_k(T) = \begin{cases} \frac{T - T_{k-1}}{\mathcal{N}_{k,\mathrm{i}}}, & r_k = 0 \\ \frac{\log \mathcal{N}_{k,\mathrm{i}} - \log \mathcal{N}(T)}{r_k}, & b_k = 0, r_k \neq 0 \\ \frac{\mathcal{N}(T)^{-b_k} - \mathcal{N}_{k,\mathrm{i}}^{-b_k}}{b_k r_k}, & b_k \neq 0, r_k \neq 0 \end{cases} . \tag{A.8}$$

Notice that the third expression above is also true for exponential growth/decline ($b_k = 1$ and $r_k \neq 0$).

A.6 Evaluation of ϕ_j^k for non-linear non-exponential generalized decline epochs

Generally, under arbitrary population size function $N(T)$, the quantity

$$\phi_j^k = \mathrm{e}^{-\binom{j}{2}\Lambda(T_{k-1})} \int_{T_{k-1}}^{T_k} \mathrm{e}^{-\binom{j}{2}\lambda_k(T)}\, dT, \tag{A.9}$$

where $\Lambda(T) = \int_0^T d\sigma/\mathcal{N}(\sigma)$, $\lambda_k(T) = \int_{T_{k-1}}^{T} d\sigma/\mathcal{N}(\sigma)$ and $\mathcal{N}(\sigma) = 2N(\sigma)$.

For generalized decline epochs ($r_k < 0$ and $b_k \notin \{0, 1\}$), in which case we didn't find feasible closed-form expression for evaluating ϕ_j^k, this quantity can be expressed in the following way:

$$\phi_j^k = \frac{\mathrm{e}^{-\binom{j}{2}\Lambda(T_{k-1})}}{\binom{j}{2}} \int_0^{\frac{\binom{j}{2}}{b_k r_k}\left(\mathcal{N}_{k,\mathrm{f}}^{-b_k} - \mathcal{N}_{k,\mathrm{i}}^{-b_k}\right)} \left(\frac{b_k r_k}{\binom{j}{2}} y + \mathcal{N}_{k,\mathrm{i}}^{-b_k}\right)^{-\frac{1}{b_k}} \mathrm{e}^{-y}\, dy. \tag{A.10}$$

The integral $\int_0^{\frac{\binom{j}{2}}{b_k r_k}\left(\mathcal{N}_{k,\mathrm{f}}^{-b_k} - \mathcal{N}_{k,\mathrm{i}}^{-b_k}\right)} \left(\frac{b_k r_k}{\binom{j}{2}} y + \mathcal{N}_{k,\mathrm{i}}^{-b_k}\right)^{-\frac{1}{b_k}} \mathrm{e}^{-y}\, dy$ is in the form of $\int_0^d (ax + b)^c e^{-x}\, dx$ where a, b, c, d are constants. We numerically evaluate this

integral by Gauss-Legendre quadrature [66]. The basic idea of Gauss-Legendre quadrature is to approximate the integrated function $f(x) = (ax + b)^c e^{-x}$ by a polynomial function of degree n, and evaluate $f(x)$ at n different points in the range $[0, d]$. The error term is $\frac{d^{(2n+1)} n!^4}{(2n+1)(2n)!^3} f^{2n}(\xi)$ [66], where $0 < \xi < d$ and $f^{(2n)}$ is the $(2n)^{\text{th}}$ derivative of f with respect to x. We choose the polynomial degree n to be 512 in this work.

A.7 Libraries used/adapted in this study

For the computation of functions $\mathcal{U}(b, x)$ and $\mathcal{M}(b, x)$, we adapted the C++ codes for the evaluation of confluent hypergeometric functions from GSL scientific library. In addition, we used the library from the link http://www.holoborodko.com/pavel/numerical-methods/numerical-integration/, which is provided by Pavel Holoborodko for Gauss-Legendre quadrature. The authors are grateful to the providers of these libraries, which are essential in the implementation of the EGGS software.

A.8 Details of simulation parameters in the second section of Results

When simulating the sequences, we used mutation rate $\mu = 1.2 \times 10^{-8}$ per base pair per generation [75] and recombination rate $\rho = 1.0 \times 10^{-8}$ per base pair per generation. To determine the amount of data for simulation, we used the number of exomes given in [127], which is about 2,500 and assumed that each

exome has 20,000 base pairs on average. To stress more the effect of linkage disequilibrium (LD) between the alleles in each exome, we decreased the number of independent loci to 1,000 and increased the length of each locus to 50,000, while keeping the total number of base pairs the same. To reduce noise in the simulated data and increase computation speed, we only kept the first 100 entries of the folded SFS and calculated the aggregate sum of the rest entries, such that there are 101 entries in total.

A.9 Details of bootstrapping

We used 200 bootstraps to obtain 95% confidence interval of the inferred parameters. For simulation studies, we randomly choose 1,000 loci from the simulated 1,000 independent loci with replacement in each bootstrap. For inference based on ESP data [41, 127], we split the sequences into 500kb regions based on SNP positions, which resulted in 882 different regions, similar to the number of loci in simulation studies. In the same manner, we then chose 882 regions with replacement for each bootstrap.

A.10 Subsampling approach

For ESP data, the successful genotype counts vary across different segregating sites. We applied the subsampling approach similarly considered in [42, 44]. For a site with n successful genotype counts, suppose there are j minor alleles and $(n-j)$ major alleles, the probability that it is of x minor alleles when subsampled

to m chromosomes is

$$\mathbb{P}[x \leftarrow m] = \frac{\binom{j}{x}\binom{n-j}{m-x}}{\binom{n}{m}} + \frac{\binom{j}{m-x}\binom{n-j}{x}}{\binom{n}{m}} \tag{A.11}$$

where $x = 0, 1, 2, \cdots, \lfloor \frac{m}{2} \rfloor$. In this work, we choose m (the number of chromosomes to subsample to) to be 7,740, which is 90% of the total number of chromosomes (8,600).

A.11 Composite log likelihood

In order to determine the fitness of a model Θ to the observed folded allele frequency counts $\boldsymbol{\mathcal{C}}$, we compute the log likelihood of the model according to

$$\mathbb{L}[\Theta] = \log \mathbb{P}[\boldsymbol{\mathcal{C}} \mid \Theta] = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \mathcal{C}_i \log \mathbb{E}[\eta_i \mid \Theta], \tag{A.12}$$

where $\mathbb{E}[\boldsymbol{\eta}|\Theta] = \left(\mathbb{E}\left[\eta_1 \mid \Theta\right], \mathbb{E}\left[\eta_2 \mid \Theta\right], \cdots, \mathbb{E}\left[\eta_{\lfloor \frac{n}{2} \rfloor} \mid \Theta\right]\right)$ is the expected folded SFS given model Θ. In this work, we considered SFS binning from the 101$^{\text{st}}$ entry to reduce the noise in later parts of the SFS: $\mathbb{E}\left[\widetilde{\boldsymbol{\eta}} \mid \Theta\right] = \left(\mathbb{E}\left[\eta_1 \mid \Theta\right], \mathbb{E}[\eta_2 \mid \Theta], \cdots, \mathbb{E}\left[\eta_{100} \mid \Theta\right], \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \mathbb{E}\left[\eta_i \mid \Theta\right]\right)$, and correspondingly the binned allele frequency counts from the data $\widetilde{\boldsymbol{\mathcal{C}}} = \left(\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_{100}, \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \mathcal{C}_i\right)$. The log likelihood after binning is computed as

$$\mathbb{L}[\Theta] = \log \mathbb{P}[\widetilde{\boldsymbol{\mathcal{C}}} \mid \Theta] = \sum_{i=1}^{101} \widetilde{\mathcal{C}}_i \log \mathbb{E}[\widetilde{\eta}_i \mid \Theta]. \tag{A.13}$$

A.12 Goodness of fit measures

In order to test how well a model SFS fits the observed data, we performed χ^2 goodness of fit test. In specific, if the observed allele frequency counts is $\boldsymbol{\mathcal{C}} =$

$(\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_d)$ (which indicates that the total number of observed segregating sites is $|\boldsymbol{\mathcal{C}}|$) and the SFS under the model is $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_d)$, then the statistic

$$\chi^2 = \sum_{i=1}^{d} \frac{(\mathcal{C}_i - |\boldsymbol{\mathcal{C}}|\xi_i)^2}{|\boldsymbol{\mathcal{C}}|\xi_i}. \tag{A.14}$$

The degree of freedom is $(d-1)$, where d is the dimension of the vector $\boldsymbol{\mathcal{C}}$. A p-value > 0.05 means we fail to reject the null hypothesis that the observed data SFS is consistent with the model SFS.

We also used another measure, Kullback-Leibler divergence or KL divergence [77], which provides a single number to relatively compare the goodness of fit between different models:

$$\mathcal{D}_{\mathrm{KL}}\left(\frac{\boldsymbol{\mathcal{C}}}{|\boldsymbol{\mathcal{C}}|}\middle\|\boldsymbol{\xi}\right) = \sum_{i=1}^{d} \frac{\mathcal{C}_i}{|\boldsymbol{\mathcal{C}}|}\left(\log\frac{\mathcal{C}_i}{|\boldsymbol{\mathcal{C}}|} - \log\xi_i\right). \tag{A.15}$$

A smaller $\mathcal{D}_{\mathrm{KL}}$ means a higher consistency between the observed and the model. The advantage of KL divergence over log likelihood is that KL divergence is a normalized measure unaffected by the total number of observed segregating sites $|\boldsymbol{\mathcal{C}}|$.

The p-values from χ^2 goodness of fit test and the KL divergence between the observed ESP data and the SFS from each of the inferred models are shown in Table S3.

A.13 Potential effect of multi-merger and simultaneous-merger events on the SFS

As sample size increases, the probability of multi-merger and simultaneous-merger events will rise, which violates the assumptions of Kingman's coales-

cent and might affect the SFS [6]. To test this effect, we used the discrete-time Wright-Fisher (DTWF) model software [6] to compute the SFS under the generalized Gazave *et al*. model with a sample size of 7,740. To shorten the computation time, we used a hybrid of DTWF and Kingman's coalescent with a time cutoff of $t_c = 212$ generations. However, it is still computationally burdensome to evaluate all 7,739 entries of the *unnormalized* SFS $\boldsymbol{\Xi}^{\text{DTWF}} = \left(\Xi_1^{\text{DTWF}}, \Xi_2^{\text{DTWF}}, \cdots, \Xi_{7739}^{\text{DTWF}}\right)$, which is needed to compute the *normalized* SFS $\boldsymbol{\xi}^{\text{DTWF}} = \frac{\boldsymbol{\Xi}^{\text{DTWF}}}{\left|\boldsymbol{\Xi}^{\text{DTWF}}\right|} = \left(\xi_1^{\text{DTWF}}, \xi_2^{\text{DTWF}}, \cdots, \xi_{7739}^{\text{DTWF}}\right)$ as is used in the inference work. We instead only evaluated the first 100 entries of $\boldsymbol{\Xi}^{\text{DTWF}}$, $\left(\Xi_1^{\text{DTWF}}, \Xi_2^{\text{DTWF}}, \cdots, \Xi_{100}^{\text{DTWF}}\right)$.

We first compared the *partially normalized* SFS under DTWF model $\boldsymbol{\xi}_{\text{partial}}^{\text{DTWF}} = \frac{1}{\sum_{i=1}^{100} \Xi_i} \left(\Xi_1^{\text{DTWF}}, \Xi_2^{\text{DTWF}}, \cdots, \Xi_{100}^{\text{DTWF}}\right)$ with the *partially normalized* SFS under Kingman's coalescent $\boldsymbol{\xi}_{\text{partial}}^{\text{Kingman}} = \frac{1}{\sum_{i=1}^{100} \xi_i^{\text{Kingman}}} \left(\xi_1^{\text{Kingman}}, \xi_2^{\text{Kingman}}, \cdots, \xi_{100}^{\text{Kingman}}\right)$, which was computed by EGGS. The two *partially normalized* SFS are very similar (Figure S9(A)). We next compared the ratio of any entry to singletons under DTWF model and Kingman's coalescent,

$$\rho_i^{\text{DTWF}} = \frac{\Xi_i^{\text{DTWF}}}{\Xi_1^{\text{DTWF}}}; \rho_i^{\text{Kingman}} = \frac{\xi_i^{\text{Kingman}}}{\xi_1^{\text{Kingman}}}, \tag{A.16}$$

where $i = 1, 2, \cdots, 100$ and we calculated the relative error,

$$\epsilon(i) = \frac{\rho_i^{\text{DTWF}} - \rho_i^{\text{Kingman}}}{\rho_i^{\text{Kingman}}} \times 100\%, \tag{A.17}$$

where $i = 1, 2, \cdots, 100$. The relative error is always less than 1% for the first 100 entries and asymptotically increases to 1% (Figure S9(B)). We then used 1% as the relative error for the rest of the SFS entries to predict the *full normalized* SFS under DTWF model. This predicted folded SFS is very similar to the folded

SFS under Kingman's coalescent (KL divergence $= 6.14 \times 10^{-6}$) and fits almost equally well to the data (KL divergence between the predicted SFS and ESP data $= 1.24 \times 10^{-4}$; p-value from χ^2 goodness of fit test $= 1$).

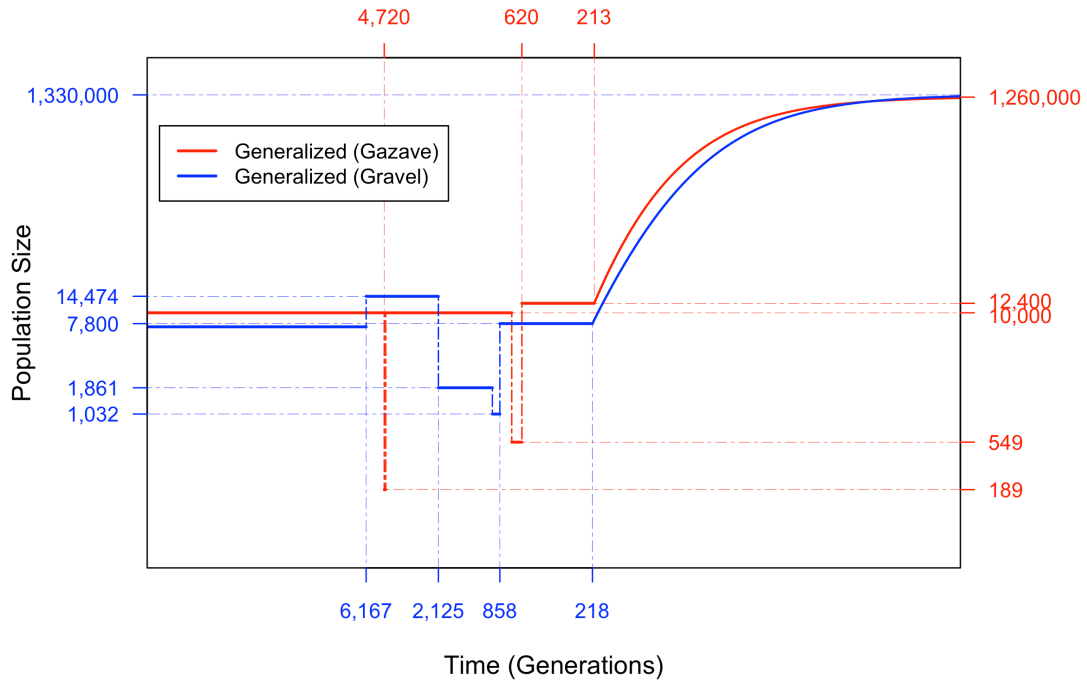

Figure A.2: **The best-fit generalized models for ESP data assuming the ancient demography in Gazave *et al.* (2014) (red) and in Gravel *et al.* (2011) (blue).** The demographic history was fixed before 620 generations ago for Gravel model and 858 generations ago for Gravel model. Both x-axis and y-axis are on log scale.

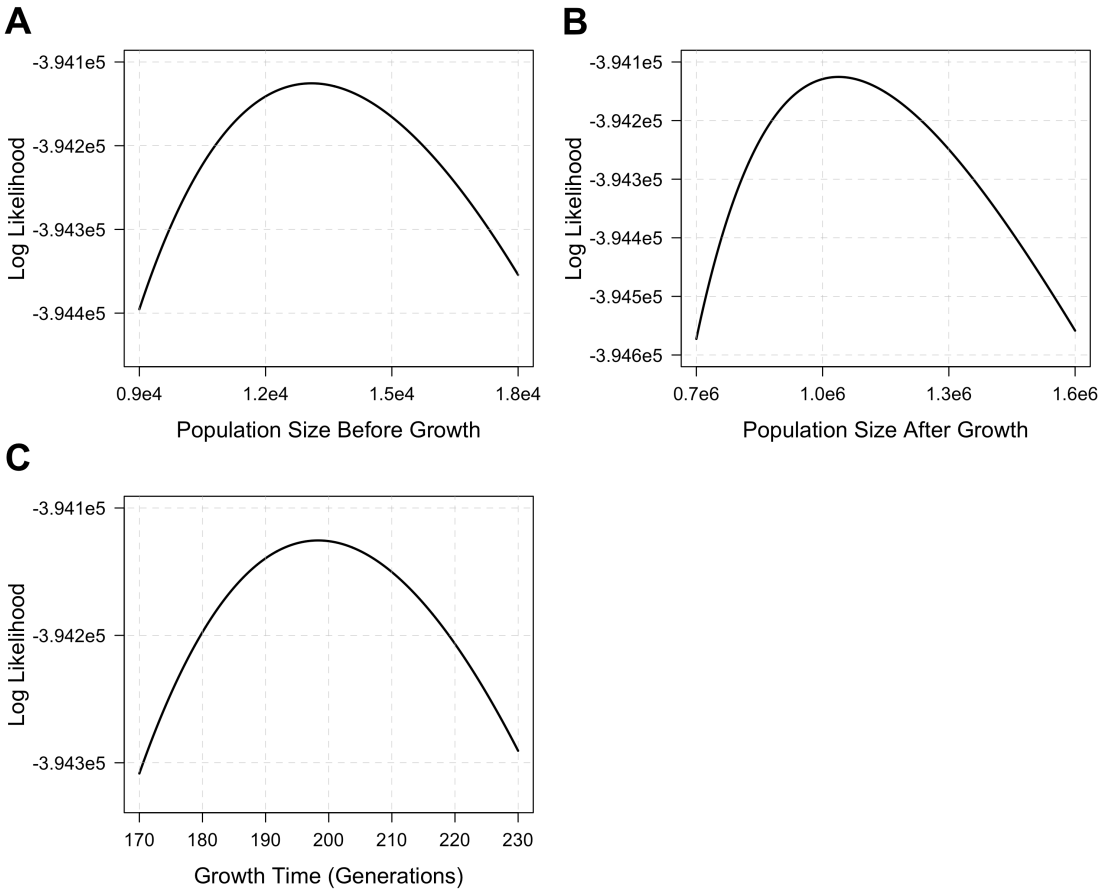

Figure A.3: **The one-dimensional log likelihood surface around the best estimates of the ESP synonymous data using exponential growth model.** (A) varying population size before growth while keeping all other parameters at corresponding best estimates; (B) varying population size after growth only; (C) varying growth time only.

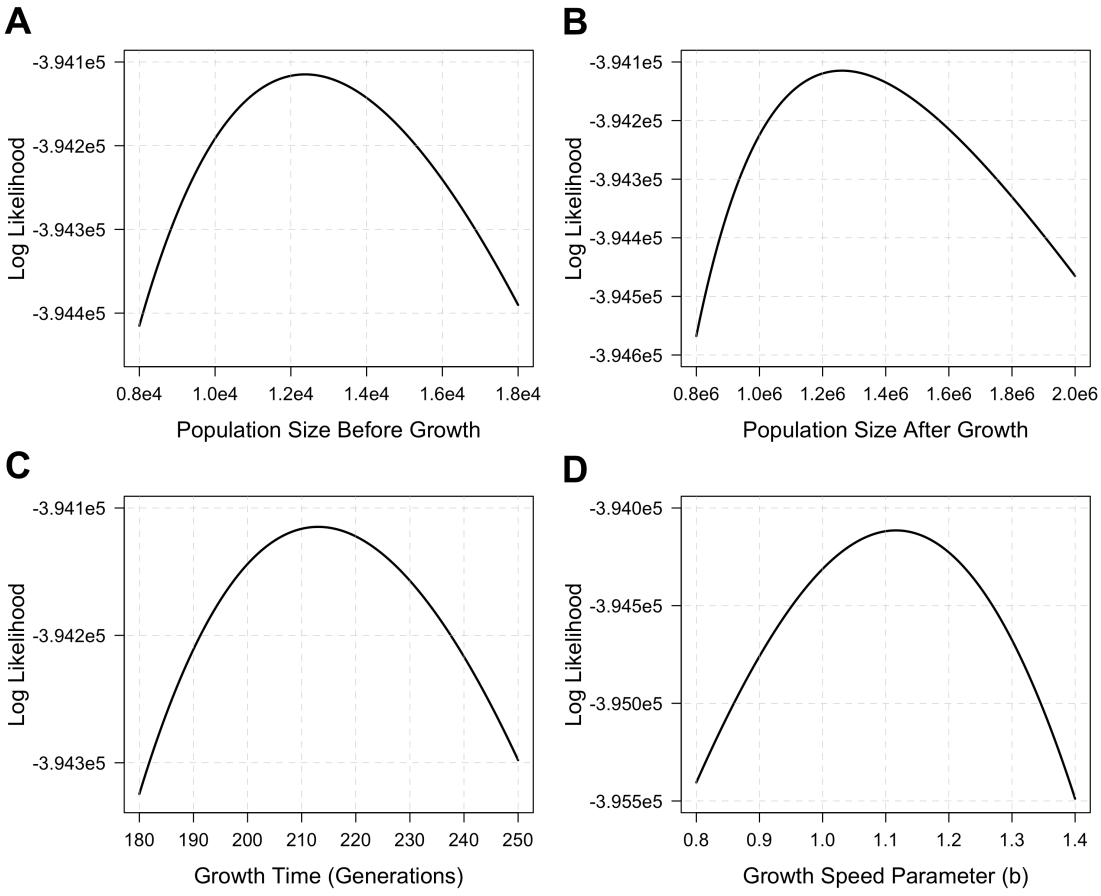

Figure A.4: **The one-dimensional log likelihood surface around the best estimates of the ESP synonymous data using generalized growth model.** (A) varying population size before growth while keeping all other parameters at corresponding best estimates; (B) varying population size after growth only; (C) varying growth time only; (D) varying growth speed parameter only.

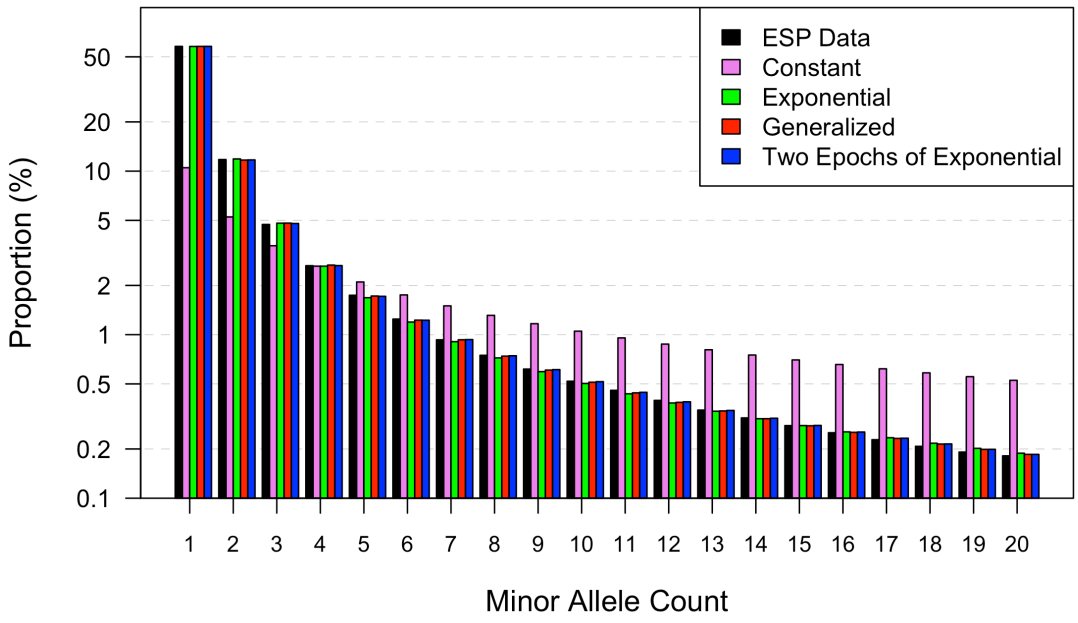

Figure A.5: **The first 20 entries of the site frequency spectra for ESP data and the inferred demographic models assuming the ancient demography in Gazave *et al.* (2014).** The SFS from the ESP data, the exponential model, the generalized growth model and the two-epoch exponential model are shown in black, green, red and blue respectively. For comparison purposes, we also included the SFS from a base model, which has a constant population size throughout history (in pink).

BIBLIOGRAPHY

[1] ABDOU, A. G., MARAEE, A. H., AND REYAD, W. Immunohistochemical expression of heat shock protein 70 in vitiligo. *Ann Diagn Pathol 17*, 3 (2013), 245–249.

[2] AHMETI, K. B., AJROUD-DRISS, S., AL-CHALABI, A., ANDERSEN, P. M., ARMSTRONG, J., BIRVE, A., BLAUW, H. M., BROWN, R. H., BRUIJN, L., CHEN, W., CHIO, A., COMEAU, M. C., CRONIN, S., DIEKSTRA, F. P., SORAYA GKAZI, A., GLASS, J. D., GRAB, J. D., GROEN, E. J., HAINES, J. L., HARDIMAN, O., HELLER, S., HUANG, J., HUNG, W. Y., CONSORTIUM, I., JAWORSKI, J. M., JONES, A., KHAN, H., LANDERS, J. E., LANGEFELD, C. D., LEIGH, P. N., MARION, M. C., MCLAUGHLIN, R. L., MEININGER, V., MELKI, J., MILLER, J. W., MORA, G., PERICAK-VANCE, M. A., RAMPERSAUD, E., ROBBERECHT, W., RUSSELL, L. P., SALACHAS, F., SARIS, C. G., SHATUNOV, A., SHAW, C. E., SIDDIQUE, N., SIDDIQUE, T., SMITH, B. N., SUFIT, R., TOPP, S., TRAYNOR, B. J., VANCE, C., VAN DAMME, P., VAN DEN BERG, L. H., VAN ES, M. A., VAN VUGHT, P. W., VELDINK, J. H., YANG, Y., ZHENG, J. G., AND ALSGEN CONSORTIUM. Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1. *Neurobiol Aging 34*, 1 (2013), 357.e7–357.e19.

[3] AHN, R., DING, Y. C., MURRAY, J., FASANO, A., GREEN, P. H., NEUHAUSEN, S. L., AND GARNER, C. Association analysis of the extended MHC region in celiac disease implicates multiple independent susceptibility loci. *PLoS One 7*, 5 (2012), e36926.

[4] ARBIZA, L., GOTTIPATI, S., SIEPEL, A., AND KEINAN, A. Contrasting X-linked and autosomal diversity across 14 human populations. *Am J Hum Genet 94*, 6 (2014), 827–844.

[5] BARANZINI, S. E., WANG, J., GIBSON, R. A., GALWEY, N., NAEGELIN, Y., BARKHOF, F., RADUE, E. W., LINDBERG, R. L., UITDEHAAG, B. M., JOHNSON, M. R., ANGELAKOPOULOU, A., HALL, L., RICHARDSON, J. C., PRINJHA, R. K., GASS, A., GEURTS, J. J., KRAGT, J., SOMBEKKE, M., VRENKEN, H., QUALLEY, P., LINCOLN, R. R., GOMEZ, R., CAILLIER, S. J., GEORGE, M. F., MOUSAVI, H., GUERRERO, R., OKUDA, D. T., CREE, B. A., GREEN, A. J., WAUBANT, E., GOODIN, D. S., PELLETIER, D., MATTHEWS, P. M., HAUSER, S. L., KAPPOS, L., POLMAN, C. H., AND OKSENBERG, J. R. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Hum Mol Genet 18*, 4 (2009), 767–778.

[6] BHASKAR, A., CLARK, A. G., AND SONG, Y. S. Distortion of genealogical properties when the sample is very large. *Proc Natl Acad Sci U S A 111*, 6 (2014), 2385–2390.

[7] BHASKAR, A., AND SONG, Y. S. Descartes' Rule of Signs and the Identifiability of Population Demographic Models from Genomic Variation Data. *Annals of Statistics 42*, 6 (2014), 2469–2493.

[8] BHASKAR, A., WANG, Y. X., AND SONG, Y. S. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Res 25*, 2 (2015), 268–279.

[9] BIANCHI, I., LLEO, A., GERSHWIN, M. E., AND INVERNIZZI, P. The x chromosome and immune associated genes. *J Autoimmun 38*, 2-3 (2012), J187–J192.

[10] BIRLEA, S. A., JIN, Y., BENNETT, D. C., HERBSTMAN, D. M., WALLACE, M. R., MCCORMACK, W. T., KEMP, E. H., GAWKRODGER, D. J., WEETMAN, A. P., PICARDO, M., LEONE, G., TAIEB, A., JOUARY, T., EZZEDINE, K., VAN GEEL, N., LAMBERT, J., OVERBECK, A., FAIN, P. R., AND SPRITZ, R. A. Comprehensive association analysis of candidate genes for generalized vitiligo supports XBP1, FOXP3, and TSLP. *J Invest Dermatol 131*, 2 (2011), 371–381.

[11] BOWIE, L. J., REDDY, P. L., AND BECK, K. R. Alpha thalassemia and its impact on other clinical conditions. *Clin Lab Med 17*, 1 (1997), 97–108.

[12] BRAUTBAR, A., BARBALIC, M., CHEN, F., BELMONT, J., VIRANI, S. S., SCHERER, S., HEGELE, R. A., AND BALLANTYNE, C. M. Rare APOA5 promoter variants associated with paradoxical HDL cholesterol decrease in response to fenofibric acid therapy. *J Lipid Res 54*, 7 (2013), 1980–1987.

[13] BRAUTBAR, A., COVARRUBIAS, D., BELMONT, J., LARA-GARDUNO, F., VIRANI, S. S., JONES, P. H., LEAL, S. M., AND BALLANTYNE, C. M. Variants in the APOA5 gene region and the response to combination therapy with statins and fenofibric acid in a randomized clinical trial of individuals with mixed dyslipidemia. *Atherosclerosis 219*, 2 (2011), 737–742.

[14] BRAUTBAR, A., VIRANI, S. S., BELMONT, J., NAMBI, V., JONES, P. H., AND BALLANTYNE, C. M. LPL gene variants affect apoC-III response to combination therapy of statins and fenofibric acid in a randomized clinical trial of individuals with mixed dyslipidemia. *J Lipid Res 53*, 3 (2012), 556–560.

[15] BROWN, M. B., AND FORSYTHE, A. B. Robust tests for equality of variances. *Journal of the American Statistical Association 69*, 346 (1974), 364–367.

[16] CARDONA, F., GUARDIOLA, M., QUEIPO-ORTUÑO, M. I., MURRI, M., RIBALTA, J., AND TINAHONES, F. J. The -1131T>C SNP of the APOA5 gene modulates response to fenofibrate treatment in patients with the metabolic syndrome: A postprandial study. *Atherosclerosis 206*, 1 (2009), 148–152.

[17] CARREL, L., AND WILLARD, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature 434*, 7031 (2005), 400–404.

[18] CHANG, D., GAO, F., SLAVNEY, A., MA, L., WALDMAN, Y. Y., SAMS, A. J., BILLING-ROSS, P., MADAR, A., SPRITZ, R., AND KEINAN, A. Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. *PLoS One 9*, 12 (2014), e113684.

[19] CHANG, D., AND KEINAN, A. Predicting signatures of "synthetic associations" and "natural associations" from empirical patterns of human genetic variation. *PLoS Comput Biol 8*, 7 (2012), e1002600.

[20] CHARLESWORTH, B., MORGAN, M. T., AND CHARLESWORTH, D. The effect of deleterious mutations on neutral molecular variation. *Genetics 134*, 4 (1993), 1289–1303.

[21] CHEN, H. The joint allele frequency spectrum of multiple populations: a coalescent theory approach. *Theoretical Population Biology 81*, 2 (2012), 179–195.

[22] CHEN, H., HEY, J., AND CHEN, K. Inferring very recent population growth rate from population-scale sequencing data: using a large-sample coalescent estimator. *Mol Biol Evol 32*, 11 (2015), 2996–3011.

[23] CHIEN, K. L., LIN, Y. L., WEN, H. C., LIN H, P., YEN, C. T., LIN, S. W., AND KAO, J. T. Common sequence variant in lipoprotein lipase gene conferring triglyceride response to fibrate treatment. *Pharmacogenomics 10*, 2 (2009), 267–276.

[24] CLAYTON, D. Testing for association on the X chromosome. *Biostatistics 9*, 4 (2008), 593–600.

[25] CLAYTON, D. G. Sex chromosomes and genetic association studies. *Genome Med 1*, 11 (2009), 110.

[26] COHEN, J. C., BOERWINKLE, E., MOSLEY, T H, J., AND HOBBS, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med 354*, 12 (2006), 1264–1272.

[27] COHEN, J. C., KISS, R. S., PERTSEMLIDIS, A., MARCEL, Y. L., MCPHERSON, R., AND HOBBS, H. H. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science 305*, 5685 (2004), 869–872.

[28] COTTON, A. M., LAM, L., AFFLECK, J. G., WILSON, I. M., PENAHERRERA, M. S., MCFADDEN, D. E., KOBOR, M. S., LAM, W. L., ROBINSON, W. P., AND BROWN, C. J. Chromosome-wide DNA methylation analysis predicts human tissue-specific X inactivation. *Hum Genet 130*, 2 (2011), 187–201.

[29] COVENTRY, A., BULL-OTTERSON, L. M., LIU, X., CLARK, A. G., MAXWELL, T. J., CROSBY, J., HIXSON, J. E., REA, T. J., MUZNY, D. M., LEWIS, L. R., WHEELER, D. A., SABO, A., LUSK, C., WEISS, K. G., AKBAR, H., CREE, A., HAWES, A. C., NEWSHAM, I., VARGHESE, R. T., VILLASANA, D., GROSS, S., JOSHI, V., SANTIBANEZ, J., MORGAN, M., CHANG, K., IV, W. H., TEMPLETON, A. R., BOERWINKLE, E., GIBBS, R., AND SING, C. F. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun 1* (2010), 131.

[30] CRONIN, S., BERGER, S., DING, J., SCHYMICK, J. C., WASHECKA, N., HERNANDEZ, D. G., GREENWAY, M. J., BRADLEY, D. G., TRAYNOR, B. J., AND HARDIMAN, O. A genome-wide association study of sporadic ALS in a homogenous Irish population. *Hum Mol Genet 17*, 5 (2008), 768–774.

[31] DICKSON, S. P., WANG, K., KRANTZ, I., HAKONARSON, H., AND GOLDSTEIN, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol 8*, 1 (2010), e1000294.

[32] DISTECHE, C. M. Dosage compensation of the sex chromosomes. *Annu Rev Genet 46* (2012), 537–560.

[33] DUERR, R. H., TAYLOR, K. D., BRANT, S. R., RIOUX, J. D., SILVERBERG, M. S., DALY, M. J., STEINHART, A. H., ABRAHAM, C., REGUEIRO, M., GRIFFITHS, A., DASSOPOULOS, T., BITTON, A., YANG, H., TARGAN, S., DATTA, L. W., KISTNER, E. O., SCHUMM, L. P., LEE, A. T., GREGERSEN,

P. K., BARMADA, M. M., ROTTER, J. I., NICOLAE, D. L., AND CHO, J. H. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science 314*, 5804 (2006), 1461–1463.

[34] ELDON, B., BIRKNER, M., BLATH, J., AND FREUND, F. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics 199*, 3 (2015), 841–856.

[35] EMERY, L. S., FELSENSTEIN, J., AND AKEY, J. M. Estimators of the human effective sex ratio detect sex biases on different timescales. *Am J Hum Genet 87*, 6 (2010), 848–856.

[36] EVANS, D. M., SPENCER, C. C., POINTON, J. J., SU, Z., HARVEY, D., KOCHAN, G., OPPERMANN, U., DILTHEY, A., PIRINEN, M., STONE, M. A., APPLETON, L., MOUTSIANAS, L., LESLIE, S., WORDSWORTH, T., KENNA, T. J., KARADERI, T., THOMAS, G. P., WARD, M. M., WEISMAN, M. H., FARRAR, C., BRADBURY, L. A., DANOY, P., INMAN, R. D., MAKSYMOWYCH, W., GLADMAN, D., RAHMAN, P., SPONDYLOARTHRITIS RESEARCH CONSORTIUM OF, C., MORGAN, A., MARZO-ORTEGA, H., BOWNESS, P., GAFFNEY, K., GASTON, J. S., SMITH, M., BRUGES-ARMAS, J., COUTO, A. R., SORRENTINO, R., PALADINI, F., FERREIRA, M. A., XU, H., LIU, Y., JIANG, L., LOPEZ-LARREA, C., DIAZ-PENA, R., LOPEZ-VAZQUEZ, A., ZAYATS, T., BAND, G., BELLENGUEZ, C., BLACKBURN, H., BLACKWELL, J. M., BRAMON, E., BUMPSTEAD, S. J., CASAS, J. P., CORVIN, A., CRADDOCK, N., DELOUKAS, P., DRONOV, S., DUNCANSON, A., EDKINS, S., FREEMAN, C., GILLMAN, M., GRAY, E., GWILLIAM, R., HAMMOND, N., HUNT, S. E., JANKOWSKI, J., JAYAKUMAR, A., LANGFORD, C., LIDDLE, J., MARKUS, H. S., MATHEW, C. G., MCCANN, O. T., MCCARTHY, M. I., PALMER, C. N., PELTONEN, L., PLOMIN, R., POTTER, S. C., RAUTANEN, A., RAVINDRARAJAH, R., RICKETTS, M., SAMANI, N., SAWCER, S. J., STRANGE, A., TREMBATH, R. C., VISWANATHAN, A. C., WALLER, M., WESTON, P., WHITTAKER, P., WIDAA, S., WOOD, N. W., MCVEAN, G., REVEILLE, J. D., WORDSWORTH, B. P., BROWN, M. A., DONNELLY, P., AUSTRALO-ANGLO-AMERICAN SPONDYLOARTHRITIS CONSORTIUM, AND WELLCOME TRUST CASE CONTROL CONSORTIUM. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat Genet 43*, 8 (2011), 761–767.

[37] EXCOFFIER, L., DUPANLOUP, I., HUERTA-SANCHEZ, E., SOUSA, V. C., AND FOLL, M. Robust demographic inference from genomic and SNP data. *PLoS Genet 9*, 10 (2013), e1003905.

[38] FAY, J. C., WYCKOFF, G. J., AND WU, C. I. Positive and negative selection on the human genome. *Genetics 158*, 3 (2001), 1227–1234.

[39] FISHER, R. A. *Statistical methods for research workers*. Oliver and Boyd, Edinburgh (UK), 1925.

[40] FRAZER, K. A., MURRAY, S. S., SCHORK, N. J., AND TOPOL, E. J. Human genetic variation and its contribution to complex traits. *Nat Rev Genet 10*, 4 (2009), 241–251.

[41] FU, W., O'CONNOR, T. D., JUN, G., KANG, H. M., ABECASIS, G., LEAL, S. M., GABRIEL, S., RIEDER, M. J., ALTSHULER, D., SHENDURE, J., NICKERSON, D. A., BAMSHAD, M. J., PROJECT, N. E. S., AND AKEY, J. M. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature 493*, 7431 (2013), 216–220.

[42] GAO, F., AND KEINAN, A. High burden of private mutations due to explosive human population growth and purifying selection. *BMC Genomics 15 Suppl 4* (2014), S3.

[43] GAO, F., AND KEINAN, A. Explosive genetic evidence for explosive human population growth. *Curr Opin Genet Dev 41* (2016), 130–139.

[44] GAZAVE, E., MA, L., CHANG, D., COVENTRY, A., GAO, F., MUZNY, D., BOERWINKLE, E., GIBBS, R. A., SING, C. F., CLARK, A. G., AND KEINAN, A. Neutral genomic regions refine models of recent rapid human population growth. *Proc Natl Acad Sci U S A 111*, 2 (2014), 757–762.

[45] GIBSON, G. Rare and common variants: twenty arguments. *Nat Rev Genet 13*, 2 (2012), 135–145.

[46] GILKS, W. P., ABBOTT, J. K., AND MORROW, E. H. Sex differences in disease genetics: evidence, evolution, and detection. *Trends Genet 30*, 10 (2014), 453–463.

[47] GOTTIPATI, S., ARBIZA, L., SIEPEL, A., CLARK, A. G., AND KEINAN, A. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nat Genet 43*, 8 (2011), 741–743.

[48] GRADSHTEĬN, I. S., RYZHIK, I. M., AND JEFFREY, A. *Table of integrals, series, and products*, 7th ed. Academic Press, Amsterdam; Boston, 2007.

[49] GRAVEL, S., HENN, B. M., GUTENKUNST, R. N., INDAP, A. R., MARTH, G. T., CLARK, A. G., YU, F., GIBBS, R. A., GENOMES, P., AND BUSTAMANTE, C. D. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A 108*, 29 (2011), 11983–11988.

[50] GUTENKUNST, R. N., HERNANDEZ, R. D., WILLIAMSON, S. H., AND BUSTAMANTE, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet 5*, 10 (2009), e1000695.

[51] HAMMER, M. F., MENDEZ, F. L., COX, M. P., WOERNER, A. E., AND WALL, J. D. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet 4*, 9 (2008), e1000202.

[52] HAMMER, M. F., WOERNER, A. E., MENDEZ, F. L., WATKINS, J. C., COX, M. P., AND WALL, J. D. The ratio of human x chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nat Genet 42*, 10 (2010), 830–831.

[53] HARRIS, K., AND NIELSEN, R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet 9*, 6 (2013), e1003521.

[54] HOWIE, B., FUCHSBERGER, C., STEPHENS, M., MARCHINI, J., AND ABECASIS, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet 44*, 8 (2012), 955–959.

[55] HUDSON, R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics 18*, 2 (2002), 337–338.

[56] HUYNH, K. D., FISCHLE, W., VERDIN, E., AND BARDWELL, V. J. BCoR, a novel corepressor involved in BCL-6 repression. *Genes Dev 14*, 14 (2000), 1810–1823.

[57] INTERNATIONAL MULTIPLE SCLEROSIS GENETICS CONSORTIUM, WELLCOME TRUST CASE CONTROL CONSORTIUM, SAWCER, S., HELLENTHAL, G., PIRINEN, M., SPENCER, C. C., PATSOPOULOS, N. A., MOUTSIANAS, L., DILTHEY, A., SU, Z., FREEMAN, C., HUNT, S. E., EDKINS, S., GRAY, E., BOOTH, D. R., POTTER, S. C., GORIS, A., BAND, G., OTURAI, A. B., STRANGE, A., SAARELA, J., BELLENGUEZ, C., FONTAINE, B., GILLMAN, M., HEMMER, B., GWILLIAM, R., ZIPP, F., JAYAKUMAR, A., MARTIN, R., LESLIE, S., HAWKINS, S., GIANNOULATOU, E., D'ALFONSO, S., BLACKBURN, H., MARTINELLI BONESCHI, F., LIDDLE, J., HARBO, H. F., PEREZ,

M. L., SPURKLAND, A., WALLER, M. J., MYCKO, M. P., RICKETTS, M., COMABELLA, M., HAMMOND, N., KOCKUM, I., MCCANN, O. T., BAN, M., WHITTAKER, P., KEMPPINEN, A., WESTON, P., HAWKINS, C., WIDAA, S., ZAJICEK, J., DRONOV, S., ROBERTSON, N., BUMPSTEAD, S. J., BARCELLOS, L. F., RAVINDRARAJAH, R., ABRAHAM, R., ALFREDSSON, L., ARDLIE, K., AUBIN, C., BAKER, A., BAKER, K., BARANZINI, S. E., BERGAMASCHI, L., BERGAMASCHI, R., BERNSTEIN, A., BERTHELE, A., BOGGILD, M., BRADFIELD, J. P., BRASSAT, D., BROADLEY, S. A., BUCK, D., BUTZKUEVEN, H., CAPRA, R., CARROLL, W. M., CAVALLA, P., CELIUS, E. G., CEPOK, S., CHIAVACCI, R., CLERGET-DARPOUX, F., CLYSTERS, K., COMI, G., COSSBURN, M., COURNU-REBEIX, I., COX, M. B., COZEN, W., CREE, B. A., CROSS, A. H., CUSI, D., DALY, M. J., DAVIS, E., DE BAKKER, P. I., DEBOUVERIE, M., D'HOOGHE M, B., DIXON, K., DOBOSI, R., DUBOIS, B., ELLINGHAUS, D., ELOVAARA, I., ESPOSITO, F., FONTENILLE, C., FOOTE, S., FRANKE, A., GALIMBERTI, D., GHEZZI, A., GLESSNER, J., GOMEZ, R., GOUT, O., GRAHAM, C., GRANT, S. F., GUERINI, F. R., HAKONARSON, H., HALL, P., HAMSTEN, A., HARTUNG, H. P., HEARD, R. N., HEATH, S., HOBART, J., HOSHI, M., INFANTE-DUARTE, C., INGRAM, G., INGRAM, W., ISLAM, T., JAGODIC, M., KABESCH, M., KERMODE, A. G., KILPATRICK, T. J., KIM, C., KLOPP, N., KOIVISTO, K., LARSSON, M., LATHROP, M., LECHNER-SCOTT, J. S., LEONE, M. A., LEPPA, V., LILJEDAHL, U., BOMFIM, I. L., LINCOLN, R. R., LINK, J., LIU, J., LORENTZEN, A. R., LUPOLI, S., MACCIARDI, F., MACK, T., MARRIOTT, M., MARTINELLI, V., MASON, D., MCCAULEY, J. L., MENTCH, F., MERO, I. L., MIHALOVA, T., MONTALBAN, X., MOTTERSHEAD, J., MYHR, K. M., NALDI, P., OLLIER, W., PAGE, A., PALOTIE, A., PELLETIER, J., PICCIO, L., PICKERSGILL, T., PIEHL, F., POBYWAJLO, S., QUACH, H. L., RAMSAY, P. P., REUNANEN, M., REYNOLDS, R., RIOUX, J. D., RODEGHER, M., ROESNER, S., RUBIO, J. P., RUCKERT, I. M., SALVETTI, M., SALVI, E., SANTANIELLO, A., SCHAEFER, C. A., SCHREIBER, S., SCHULZE, C., SCOTT, R. J., SELLEBJERG, F., SELMAJ, K. W., SEXTON, D., SHEN, L., SIMMS-ACUNA, B., SKIDMORE, S., SLEIMAN, P. M., SMESTAD, C., SORENSEN, P. S., SONDERGAARD, H. B., STANKOVICH, J., STRANGE, R. C., SULONEN, A. M., SUNDQVIST, E., SYVANEN, A. C., TADDEO, F., TAYLOR, B., BLACKWELL, J. M., TIENARI, P., BRAMON, E., TOURBAH, A., BROWN, M. A., TRONCZYNSKA, E., CASAS, J. P., TUBRIDY, N., CORVIN, A., VICKERY, J., JANKOWSKI, J., VILLOSLADA, P., MARKUS, H. S., WANG, K., MATHEW, C. G., WASON, J., PALMER, C. N., WICHMANN, H. E., PLOMIN, R., WILLOUGHBY, E., RAUTANEN, A., WINKELMANN, J., WITTIG, M., TREMBATH, R. C., YAOUANQ, J., VISWANATHAN, A. C., ZHANG, H., WOOD, N. W., ZUVICH, R., DELOUKAS, P., LANGFORD, C., DUNCANSON, A., OKSENBERG, J. R., PERICAK-VANCE, M. A., HAINES, J. L., OLSSON, T., HILLERT, J., IVINSON, A. J., DE JAGER, P. L., PELTONEN, L., STEWART,

G. J., HAFLER, D. A., HAUSER, S. L., MCVEAN, G., DONNELLY, P., AND COMPSTON, A. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature 476*, 7359 (2011), 214–219.

[58] JIANG, B., ZHANG, X., ZUO, Y., AND KANG, G. A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. *J Theor Biol 277*, 1 (2011), 67–73.

[59] JIN, X., CHEN, Y. P., CHEN, S. H., AND XIANG, Z. Association between Helicobacter Pylori infection and ulcerative colitis–a case control study from China. *Int J Med Sci 10*, 11 (2013), 1479–1484.

[60] JIN, Y., BIRLEA, S. A., FAIN, P. R., FERRARA, T. M., BEN, S., RICCARDI, S. L., COLE, J. B., GOWAN, K., HOLLAND, P. J., BENNETT, D. C., LUITEN, R. M., WOLKERSTORFER, A., VAN DER VEEN, J. P., HARTMANN, A., EICHNER, S., SCHULER, G., VAN GEEL, N., LAMBERT, J., KEMP, E. H., GAWKRODGER, D. J., WEETMAN, A. P., TAIEB, A., JOUARY, T., EZZEDINE, K., WALLACE, M. R., MCCORMACK, W. T., PICARDO, M., LEONE, G., OVERBECK, A., SILVERBERG, N. B., AND SPRITZ, R. A. Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nat Genet 44*, 6 (2012), 676–680.

[61] JIN, Y., BIRLEA, S. A., FAIN, P. R., GOWAN, K., RICCARDI, S. L., HOLLAND, P. J., MAILLOUX, C. M., SUFIT, A. J., HUTTON, S. M., AMADI-MYERS, A., BENNETT, D. C., WALLACE, M. R., MCCORMACK, W. T., KEMP, E. H., GAWKRODGER, D. J., WEETMAN, A. P., PICARDO, M., LEONE, G., TAIEB, A., JOUARY, T., EZZEDINE, K., VAN GEEL, N., LAMBERT, J., OVERBECK, A., AND SPRITZ, R. A. Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo. *N Engl J Med 362*, 18 (2010), 1686–1697.

[62] JOHANSEN, C. T., AND HEGELE, R. A. Genetic bases of hypertriglyceridemic phenotypes. *Curr Opin Lipidol 22*, 4 (2011), 247–253.

[63] JONES, P. H., BAYS, H. E., DAVIDSON, M. H., KELLY, M. T., BUTTLER, S. M., SETZE, C. M., SLEEP, D. J., AND STOLZENBACH, J. C. Evaluation of a new formulation of fenofibric acid, ABT-335, co-administered with statins: study design and rationale of a phase III clinical programme. *Clin Drug Investig 28*, 10 (2008), 625–634.

[64] JONES, P. H., DAVIDSON, M. H., KASHYAP, M. L., KELLY, M. T., BUTTLER, S. M., SETZE, C. M., SLEEP, D. J., AND STOLZENBACH, J. C. Efficacy and safety of ABT-335 (fenofibric acid) in combination with rosuvas-

tatin in patients with mixed dyslipidemia: a phase 3 study. *Atherosclerosis 204*, 1 (2009), 208–215.

[65] JU, T., AND CUMMINGS, R. D. Protein glycosylation: chaperone mutation in Tn syndrome. *Nature 437*, 7063 (2005), 1252.

[66] KAHANER, D., MOLER, C. B., NASH, S., AND FORSYTHE, G. E. *Numerical methods and software*. Prentice Hall, Englewood Cliffs, N.J., 1988.

[67] KEINAN, A., AND CLARK, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science 336*, 6082 (2012), 740–743.

[68] KEINAN, A., MULLIKIN, J. C., PATTERSON, N., AND REICH, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet 39*, 10 (2007), 1251–1255.

[69] KEINAN, A., MULLIKIN, J. C., PATTERSON, N., AND REICH, D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat Genet 41*, 1 (2009), 66–70.

[70] KEINAN, A., AND REICH, D. Can a sex-biased human demography account for the reduced effective population size of chromosome x in non-africans? *Mol Biol Evol 27*, 10 (2010), 2312–2321.

[71] KIEZUN, A., GARIMELLA, K., DO, R., STITZIEL, N. O., NEALE, B. M., MCLAREN, P. J., GUPTA, N., SKLAR, P., SULLIVAN, P. F., MORAN, J. L., HULTMAN, C. M., LICHTENSTEIN, P., MAGNUSSON, P., LEHNER, T., SHUGART, Y. Y., PRICE, A. L., DE BAKKER, P. I., PURCELL, S. M., AND SUNYAEV, S. Exome sequencing and the genetic basis of complex traits. *Nat Genet 44*, 6 (2012), 623–630.

[72] KIMURA, M., AND OHTA, T. The age of a neutral mutant persisting in a finite population. *Genetics 75*, 1 (1973), 199–212.

[73] KINGMAN, K. F. C. On the genealogy of large populations. *J Appl Probab 19* (1982), 27–43.

[74] KINGMAN, K. F. C. The coalescent. *Stochastic Process Appl 13*, 3 (1982), 235–248.

[75] KONG, A., FRIGGE, M. L., MASSON, G., BESENBACHER, S., SULEM, P., MAGNUSSON, G., GUDJONSSON, S. A., SIGURDSSON, A., JONASDOTTIR, A., JONASDOTTIR, A., WONG, W. S., SIGURDSSON, G., WALTERS, G. B., STEINBERG, S., HELGASON, H., THORLEIFSSON, G., GUDBJARTSSON, D. F., HELGASON, A., MAGNUSSON, O. T., THORSTEINSDOTTIR, U., AND STEFANSSON, K. Rate of de novo mutations and the importance of father's age to disease risk. *Nature 488*, 7412 (2012), 471–475.

[76] KORN, J. M., KURUVILLA, F. G., MCCARROLL, S. A., WYSOKER, A., NEMESH, J., CAWLEY, S., HUBBELL, E., VEITCH, J., COLLINS, P. J., DARVISHI, K., LEE, C., NIZZARI, M. M., GABRIEL, S. B., PURCELL, S., DALY, M. J., AND ALTSHULER, D. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet 40*, 10 (2008), 1253–1260.

[77] KULLBACK, S., AND LEIBLER, R. A. On information and sufficiency. *Ann Math Stat 22*, 1 (1951), 79–86.

[78] LAAKSOVIRTA, H., PEURALINNA, T., SCHYMICK, J. C., SCHOLZ, S. W., LAI, S. L., MYLLYKANGAS, L., SULKAVA, R., JANSSON, L., HERNANDEZ, D. G., GIBBS, J. R., NALLS, M. A., HECKERMAN, D., TIENARI, P. J., AND TRAYNOR, B. J. Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. *Lancet Neurol 9*, 10 (2010), 978–985.

[79] LAI, C. Q., ARNETT, D. K., CORELLA, D., STRAKA, R. J., TSAI, M. Y., PEACOCK, J. M., ADICONIS, X., PARNELL, L. D., HIXSON, J. E., PROVINCE, M. A., AND ORDOVAS, J. M. Fenofibrate effect on triglyceride and postprandial response of Apolipoprotein A5 variants: the GOLDN study. *Arterioscler Thromb Vasc Biol 27*, 6 (2007), 1417–1425.

[80] LAMBERT, C. A., CONNELLY, C. F., MADEOY, J., QIU, R., OLSON, M., AND AKEY, J. M. Highly punctuated patterns of population structure on the X chromosome and implications for African evolutionary history. *Am J Hum Genet 86*, 1 (2010), 34–44.

[81] LERNER, D. J., AND KANNEL, W. B. Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the Framingham population. *Am Heart J 111*, 2 (1986), 383–390.

[82] LI, B., AND LEAL, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet 83*, 3 (2008), 311–321.

[83] LI, H., AND DURBIN, R. Inference of human population history from individual whole-genome sequences. *Nature 475*, 7357 (2011), 493–496.

[84] LI, Y. R., LI, J., ZHAO, S. D., BRADFIELD, J. P., MENTCH, F. D., MAGGADOTTIR, S. M., HOU, C., ABRAMS, D. J., CHANG, D., GAO, F., GUO, Y., WEI, Z., CONNOLLY, J. J., CARDINALE, C. J., BAKAY, M., GLESSNER, J. T., LI, D., KAO, C., THOMAS, K. A., QIU, H., CHIAVACCI, R. M., KIM, C. E., WANG, F., SNYDER, J., RICHIE, M. D., FLATO, B., FORRE, O., DENSON, L. A., THOMPSON, S. D., BECKER, M. L., GUTHERY, S. L., LATIANO, A., PEREZ, E., RESNICK, E., RUSSELL, R. K., WILSON, D. C., SILVERBERG, M. S., ANNESE, V., LIE, B. A., PUNARO, M., DUBINSKY, M. C., MONOS, D. S., STRISCIUGLIO, C., STAIANO, A., MIELE, E., KUGATHASAN, S., ELLIS, J. A., MUNRO, J. E., SULLIVAN, K. E., WISE, C. A., CHAPEL, H., CUNNINGHAM-RUNDLES, C., GRANT, S. F., ORANGE, J. S., SLEIMAN, P. M., BEHRENS, E. M., GRIFFITHS, A. M., SATSANGI, J., FINKEL, T. H., KEINAN, A., PRAK, E. T., POLYCHRONAKOS, C., BALDASSANO, R. N., LI, H., KEATING, B. J., AND HAKONARSON, H. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nat Med 21*, 9 (2015), 1018–1027.

[85] LIBERT, C., DEJAGER, L., AND PINHEIRO, I. The X chromosome in immune functions: when a chromosome makes the difference. *Nat Rev Immunol 10*, 8 (2010), 594–604.

[86] LIU, J. Z., MCRAE, A. F., NYHOLT, D. R., MEDLAND, S. E., WRAY, N. R., BROWN, K. M., INVESTIGATORS, A., HAYWARD, N. K., MONTGOMERY, G. W., VISSCHER, P. M., MARTIN, N. G., AND MACGREGOR, S. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet 87*, 1 (2010), 139–145.

[87] LIU, X., AND FU, Y. X. Exploring population size changes using SNP frequency spectra. *Nat Genet 47*, 5 (2015), 555–559.

[88] LIU, Y., ORDOVAS, J. M., GAO, G., PROVINCE, M., STRAKA, R. J., TSAI, M. Y., LAI, C. Q., ZHANG, K., BORECKI, I., HIXSON, J. E., ALLISON, D. B., AND ARNETT, D. K. Pharmacogenetic association of the APOA1/C3/A4/A5 gene cluster and lipid responses to fenofibrate: the genetics of lipid-lowering drugs and diet network study. *Pharmacogenet Genomics 19*, 2 (2009), 161–169.

[89] LOHMUELLER, K. E., DEGENHARDT, J. D., AND KEINAN, A. Sex-averaged recombination and mutation rates on the X chromosome: a com-

ment on Labuda et al. *Am J Hum Genet 86*, 6 (2010), 978–980; author reply 980–981.

[90] LOLEY, C., ZIEGLER, A., AND KONIG, I. R. Association tests for X-chromosomal markers–a comparison of different test statistics. *Hum Hered 71*, 1 (2011), 23–36.

[91] LUTHER, J., DAVE, M., HIGGINS, P. D., AND KAO, J. Y. Association between Helicobacter pylori infection and inflammatory bowel disease: a meta-analysis and systematic review of the literature. *Inflamm Bowel Dis 16*, 6 (2010), 1077–1084.

[92] MA, L., BALLANTYNE, C. M., BELMONT, J. W., KEINAN, A., AND BRAUTBAR, A. Interaction between SNPs in the RXRA and near ANGPTL3 gene region inhibits apoB reduction after statin-fenofibric acid therapy in individuals with mixed dyslipidemia. *J Lipid Res 53*, 11 (2012), 2425–2428.

[93] MA, L., CLARK, A. G., AND KEINAN, A. Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genet 9*, 2 (2013), e1003321.

[94] MA, L., HOFFMAN, G., AND KEINAN, A. X-inactivation informs variance-based testing for X-linked association of a quantitative trait. *BMC Genomics 16* (2015), 241.

[95] MACLEOD, I. M., LARKIN, D. M., LEWIN, H. A., HAYES, B. J., AND GODDARD, M. E. Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. *Mol Biol Evol 30*, 9 (2013), 2209–2223.

[96] MAHER, B. Personal genomes: The case of the missing heritability. *Nature 456*, 7218 (2008), 18–21.

[97] MAILMAN, M. D., FEOLO, M., JIN, Y., KIMURA, M., TRYKA, K., BAGOUTDINOV, R., HAO, L., KIANG, A., PASCHALL, J., PHAN, L., POPOVA, N., PRETEL, S., ZIYABARI, L., LEE, M., SHAO, Y., WANG, Z. Y., SIROTKIN, K., WARD, M., KHOLODOV, M., ZBICZ, K., BECK, J., KIMELMAN, M., SHEVELEV, S., PREUSS, D., YASCHENKO, E., GRAEFF, A., OSTELL, J., AND SHERRY, S. T. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet 39*, 10 (2007), 1181–1186.

[98] MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., A, H. L., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., S, W. A., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F., MCCARROLL, S. A., AND VISSCHER, P. M. Finding the missing heritability of complex diseases. *Nature 461*, 7265 (2009), 747–753.

[99] MARTH, G. T., CZABARKA, E., MURVAI, J., AND SHERRY, S. T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics 166*, 1 (2004), 351–372.

[100] MATANOSKI, G., TAO, X., ALMON, L., ADADE, A. A., AND DAVIES-COLE, J. O. Demographics and tumor characteristics of colorectal cancers in the united states, 1998-2001. *Cancer 107*, 5 Suppl (2006), 1112–1120.

[101] MENG, X. L., AND RUBIN, D. B. Maximum-likelihood-estimation via the ECM algorithm – a general framework. *Biometrika 80*, 2 (1993), 267–278.

[102] MORGENTHALER, S., AND THILLY, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res 615*, 1-2 (2007), 28–56.

[103] MORRIS, A. P., AND ZEGGINI, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol 34*, 2 (2010), 188–193.

[104] MOSENSON, J. A., EBY, J. M., HERNANDEZ, C., AND LE POOLE, I. C. A central role for inducible heat-shock protein 70 in autoimmune vitiligo. *Exp Dermatol 22*, 9 (2013), 566–9.

[105] MOSENSON, J. A., ZLOZA, A., KLARQUIST, J., BARFUSS, A. J., GUEVARA-PATINO, J. A., AND POOLE, I. C. HSP70i is a critical component of the immune response leading to vitiligo. *Pigment Cell Melanoma Res 25*, 1 (2012), 88–98.

[106] MUSCAT, J. E., RICHIE, J P, J., THOMPSON, S., AND WYNDER, E. L. Gender differences in smoking and risk for oral cancer. *Cancer Res 56*, 22 (1996), 5192–5197.

[107] NAIR, R. P., DUFFIN, K. C., HELMS, C., DING, J., STUART, P. E., GOLDGAR, D., GUDJONSSON, J. E., LI, Y., TEJASVI, T., FENG, B. J., RUETHER, A., SCHREIBER, S., WEICHENTHAL, M., GLADMAN, D., RAHMAN, P., SCHRODI, S. J., PRAHALAD, S., GUTHERY, S. L., FISCHER, J., LIAO, W., KWOK, P. Y., MENTER, A., LATHROP, G. M., WISE, C. A., BEGOVICH, A. B., VOORHEES, J. J., ELDER, J. T., KRUEGER, G. G., BOWCOCK, A. M., ABECASIS, G. R., AND COLLABORATIVE ASSOCIATION STUDY OF PSORIASIS. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat Genet 41*, 2 (2009), 199–204.

[108] NELSON, M. R., WEGMANN, D., EHM, M. G., KESSNER, D., ST JEAN, P., VERZILLI, C., SHEN, J., TANG, Z., BACANU, S. A., FRASER, D., WARREN, L., APONTE, J., ZAWISTOWSKI, M., LIU, X., ZHANG, H., ZHANG, Y., LI, J., LI, Y., LI, L., WOOLLARD, P., TOPP, S., HALL, M. D., NANGLE, K., WANG, J., ABECASIS, G., CARDON, L. R., ZOLLNER, S., WHITTAKER, J. C., CHISSOE, S. L., NOVEMBRE, J., AND MOOSER, V. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science 337*, 6090 (2012), 100–104.

[109] OBER, C., LOISEL, D. A., AND GILAD, Y. Sex-specific genetic architecture of human disease. *Nat Rev Genet 9*, 12 (2008), 911–922.

[110] POLANSKI, A., BOBROWSKI, A., AND KIMMEL, M. A note on distributions of times to coalescence, under time-dependent population size. *Theoretical Population Biology 63*, 1 (2003), 33–40.

[111] POLANSKI, A., AND KIMMEL, M. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics 165*, 1 (2003), 427–436.

[112] PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A., AND REICH, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet 38*, 8 (2006), 904–909.

[113] PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J., AND SHAM, P. C. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet 81*, 3 (2007), 559–575.

[114] QI, L., CORNELIS, M. C., KRAFT, P., STANYA, K. J., LINDA KAO, W. H., PANKOW, J. S., DUPUIS, J., FLOREZ, J. C., FOX, C. S., PARE, G., SUN, Q.,

GIRMAN, C. J., LAURIE, C. C., MIREL, D. B., MANOLIO, T. A., CHASMAN, D. I., BOERWINKLE, E., RIDKER, P. M., HUNTER, D. J., MEIGS, J. B., LEE, C. H., HU, F. B., VAN DAM, R. M., META-ANALYSIS OF GLUCOSE AND INSULIN-RELATED TRAITS CONSORTIUM (MAGIC), AND DIABETES GENETICS REPLICATION AND META-ANALYSIS (DIAGRAM) CONSORTIUM. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. *Hum Mol Genet 19*, 13 (2010), 2706–15.

[115] QUINTERO, O. L., AMADOR-PATARROYO, M. J., MONTOYA-ORTIZ, G., ROJAS-VILLARRAGA, A., AND ANAYA, J. M. Autoimmune disease and gender: plausible mechanisms for the female predominance of autoimmunity. *J Autoimmun 38*, 2–3 (2012), J109–J119.

[116] RANDALL, J. C., WINKLER, T. W., KUTALIK, Z., BERNDT, S. I., JACKSON, A. U., MONDA, K. L., KILPELAINEN, T. O., ESKO, T., MAGI, R., LI, S., WORKALEMAHU, T., FEITOSA, M. F., CROTEAU-CHONKA, D. C., DAY, F. R., FALL, T., FERREIRA, T., GUSTAFSSON, S., LOCKE, A. E., MATHIESON, I., SCHERAG, A., VEDANTAM, S., WOOD, A. R., LIANG, L., STEINTHORSDOTTIR, V., THORLEIFSSON, G., DERMITZAKIS, E. T., DIMAS, A. S., KARPE, F., MIN, J. L., NICHOLSON, G., CLEGG, D. J., PERSON, T., KROHN, J. P., BAUER, S., BUECHLER, C., EISINGER, K., CONSORTIUM, D., BONNEFOND, A., FROGUEL, P., INVESTIGATORS, M., HOTTENGA, J. J., PROKOPENKO, I., WAITE, L. L., HARRIS, T. B., SMITH, A. V., SHULDINER, A. R., MCARDLE, W. L., CAULFIELD, M. J., MUNROE, P. B., GRONBERG, H., CHEN, Y. D., LI, G., BECKMANN, J. S., JOHNSON, T., THORSTEINSDOTTIR, U., TEDER-LAVING, M., KHAW, K. T., WAREHAM, N. J., ZHAO, J. H., AMIN, N., OOSTRA, B. A., KRAJA, A. T., PROVINCE, M. A., CUPPLES, L. A., HEARD-COSTA, N. L., KAPRIO, J., RIPATTI, S., SURAKKA, I., COLLINS, F. S., SARAMIES, J., TUOMILEHTO, J., JULA, A., SALOMAA, V., ERDMANN, J., HENGSTENBERG, C., LOLEY, C., SCHUNKERT, H., LAMINA, C., WICHMANN, H. E., ALBRECHT, E., GIEGER, C., HICKS, A. A., JOHANSSON, A., PRAMSTALLER, P. P., KATHIRESAN, S., SPELIOTES, E. K., PENNINX, B., HARTIKAINEN, A. L., JARVELIN, M. R., GYLLENSTEN, U., BOOMSMA, D. I., CAMPBELL, H., WILSON, J. F., CHANOCK, S. J., FARRALL, M., GOEL, A., MEDINA-GOMEZ, C., RIVADENEIRA, F., ESTRADA, K., UITTERLINDEN, A. G., HOFMAN, A., ZILLIKENS, M. C., DEN HEIJER, M., KIEMENEY, L. A., MASCHIO, A., HALL, P., TYRER, J., TEUMER, A., VOLZKE, H., KOVACS, P., TONJES, A., MANGINO, M., SPECTOR, T. D., HAYWARD, C., RUDAN, I., HALL, A. S., SAMANI, N. J., ATTWOOD, A. P., SAMBROOK, J. G., HUNG, J., PALMER, L. J., LOKKI, M. L., SINISALO, J., BOUCHER, G., HUIKURI, H., LORENTZON, M., OHLSSON, C., EKLUND, N., ERIKSSON, J. G., BARLASSINA, C., RIVOLTA, C., NOLTE, I. M., SNIEDER, H.,

VAN DER KLAUW, M. M., VAN VLIET-OSTAPTCHOUK, J. V., GEJMAN, P. V., SHI, J., JACOBS, K. B., WANG, Z., BAKKER, S. J., MATEO LEACH, I., NAVIS, G., VAN DER HARST, P., MARTIN, N. G., MEDLAND, S. E., MONTGOMERY, G. W., YANG, J., CHASMAN, D. I., RIDKER, P. M., ROSE, L. M., LEHTIMAKI, T., RAITAKARI, O., ABSHER, D., IRIBARREN, C., BASART, H., HOVINGH, K. G., HYPPONEN, E., POWER, C., ANDERSON, D., BEILBY, J. P., HUI, J., JOLLEY, J., SAGER, H., BORNSTEIN, S. R., SCHWARZ, P. E., KRISTIANSSON, K., PEROLA, M., LINDSTROM, J., SWIFT, A. J., UUSITUPA, M., ATALAY, M., LAKKA, T. A., RAURAMAA, R., BOLTON, J. L., FOWKES, G., FRASER, R. M., PRICE, J. F., FISCHER, K., KRJUTA KOV, K., METSPALU, A., MIHAILOV, E., LANGENBERG, C., LUAN, J., ONG, K. K., CHINES, P. S., KEINANEN-KIUKAANNIEMI, S. M., SAARISTO, T. E., EDKINS, S., FRANKS, P. W., HALLMANS, G., SHUNGIN, D., MORRIS, A. D., PALMER, C. N., ERBEL, R., MOEBUS, S., NOTHEN, M. M., PECHLIVANIS, S., HVEEM, K., NARISU, N., HAMSTEN, A., HUMPHRIES, S. E., STRAWBRIDGE, R. J., TREMOLI, E., GRALLERT, H., THORAND, B., ILLIG, T., KOENIG, W., MULLER-NURASYID, M., PETERS, A., BOEHM, B. O., KLEBER, M. E., MARZ, W., WINKELMANN, B. R., KUUSISTO, J., LAAKSO, M., ARVEILER, D., CESANA, G., KUULASMAA, K., VIRTAMO, J., YARNELL, J. W., KUH, D., WONG, A., LIND, L., DE FAIRE, U., GIGANTE, B., MAGNUSSON, P. K., PEDERSEN, N. L., DEDOUSSIS, G., DIMITRIOU, M., KOLOVOU, G., KANONI, S., STIRRUPS, K., BONNYCASTLE, L. L., NJOLSTAD, I., WILSGAARD, T., GANNA, A., REHNBERG, E., HINGORANI, A., KIVIMAKI, M., KUMARI, M., ASSIMES, T. L., BARROSO, I., BOEHNKE, M., BORECKI, I. B., DELOUKAS, P., FOX, C. S., FRAYLING, T., GROOP, L. C., HARITUNIANS, T., HUNTER, D., INGELSSON, E., KAPLAN, R., MOHLKE, K. L., O'CONNELL, J. R., SCHLESSINGER, D., STRACHAN, D. P., STEFANSSON, K., VAN DUIJN, C. M., ABECASIS, G. R., MCCARTHY, M. I., HIRSCHHORN, J. N., QI, L., LOOS, R. J., LINDGREN, C. M., NORTH, K. E., AND HEID, I. M. Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet 9*, 6 (2013), e1003500.

[117] REPPELL, M., BOEHNKE, M., AND ZOLLNER, S. FTEC: a coalescent simulator for modeling faster than exponential growth. *Bioinformatics 28*, 9 (2012), 1282–1283.

[118] REPPELL, M., BOEHNKE, M., AND ZOLLNER, S. The impact of accelerating faster than exponential population growth on genetic variation. *Genetics 196*, 3 (2014), 819–828.

[119] ROACH, J. C., GLUSMAN, G., SMIT, A. F., HUFF, C. D., HUBLEY, R.,

SHANNON, P. T., ROWEN, L., PANT, K. P., GOODMAN, N., BAMSHAD, M., SHENDURE, J., DRMANAC, R., JORDE, L. B., HOOD, L., AND GALAS, D. J. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science 328*, 5978 (2010), 636–639.

[120] SCALLY, A., AND DURBIN, R. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet 13*, 10 (2012), 745–753.

[121] SCHIFFELS, S., AND DURBIN, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet 46*, 8 (2014), 919–925.

[122] SHEEHAN, S., HARRIS, K., AND SONG, Y. S. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. *Genetics 194*, 3 (2013), 647–662.

[123] STOUFFER, S. A. *The American soldier*. Studies in social psychology in World War II. Princeton University Press, Princeton, 1949.

[124] TAKAHATA, N., AND NEI, M. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics 110*, 2 (1985), 325–344.

[125] TAVARE, S. Line-of-descent and genealogical processes, and their applications in population-genetics models. *Theor Pop Biol 26*, 2 (1984), 119–164.

[126] TAYLOR, P. N., PORCU, E., CHEW, S., CAMPBELL, P. J., TRAGLIA, M., BROWN, S. J., MULLIN, B. H., SHIHAB, H. A., MIN, J., WALTER, K., MEMARI, Y., HUANG, J., BARNES, M. R., BEILBY, J. P., CHAROEN, P., DANECEK, P., DUDBRIDGE, F., FORGETTA, V., GREENWOOD, C., GRUNDBERG, E., JOHNSON, A. D., HUI, J., LIM, E. M., MCCARTHY, S., MUDDYMAN, D., PANICKER, V., PERRY, J. R., BELL, J. T., YUAN, W., RELTON, C., GAUNT, T., SCHLESSINGER, D., ABECASIS, G., CUCCA, F., SURDULESCU, G. L., WOLTERSDORF, W., ZEGGINI, E., ZHENG, H. F., TONIOLO, D., DAYAN, C. M., NAITZA, S., WALSH, J. P., SPECTOR, T., DAVEY SMITH, G., DURBIN, R., RICHARDS, J. B., SANNA, S., SORANZO, N., TIMPSON, N. J., WILSON, S. G., AND THE UK10K CONSORTIUM. Whole-genome sequence-based analysis of thyroid function. *Nat Commun 6* (2015), 5681.

[127] TENNESSEN, J. A., BIGHAM, A. W., O'CONNOR, T. D., FU, W., KENNY, E. E., GRAVEL, S., MCGEE, S., DO, R., LIU, X., JUN, G., KANG, H. M.,

JORDAN, D., LEAL, S. M., GABRIEL, S., RIEDER, M. J., ABECASIS, G., ALTSHULER, D., NICKERSON, D. A., BOERWINKLE, E., SUNYAEV, S., BUSTAMANTE, C. D., BAMSHAD, M. J., AKEY, J. M., BROAD G O, SEATTLE G O, AND NHLBI EXOME SEQUENCING PROJECT. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science 337*, 6090 (2012), 64–69.

[128] TERHORST, J., AND SONG, Y. S. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proc Natl Acad Sci U S A 112*, 25 (2015), 7677–7682.

[129] THE ACCORD STUDY GROUP, GINSBERG, H. N., ELAM, M. B., LOVATO, L. C., CROUSE, J. R., R., LEITER, L. A., LINZ, P., FRIEDEWALD, W. T., BUSE, J. B., GERSTEIN, H. C., PROBSTFIELD, J., GRIMM, R. H., ISMAIL-BEIGI, F., BIGGER, J. T., GOFF, D C, J., CUSHMAN, W. C., SIMONS-MORTON, D. G., AND BYINGTON, R. P. Effects of combination lipid therapy in type 2 diabetes mellitus. *N Engl J Med 362*, 17 (2010), 1563–1574.

[130] THORNTON, T., ZHANG, Q., CAI, X., OBER, C., AND MCPEEK, M. S. XM: association testing on the X-chromosome in case-control samples with related individuals. *Genet Epidemiol 36*, 5 (2012), 438–50.

[131] TRIGLYCERIDE CORONARY DISEASE GENETICS CONSORTIUM AND EMERGING RISK FACTORS COLLABORATION, SARWAR, N., SANDHU, M. S., RICKETTS, S. L., BUTTERWORTH, A. S., DI ANGELANTONIO, E., BOEKHOLDT, S. M., OUWEHAND, W., WATKINS, H., SAMANI, N. J., SALEHEEN, D., LAWLOR, D., REILLY, M. P., HINGORANI, A. D., TALMUD, P. J., AND DANESH, J. Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies. *Lancet 375*, 9726 (2010), 1634–1639.

[132] TRYKA, K. A., HAO, L., STURCKE, A., JIN, Y., WANG, Z. Y., ZIYABARI, L., LEE, M., POPOVA, N., SHAROPOVA, N., KIMURA, M., AND FEOLO, M. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res 42*, Database issue (2014), D975–D979.

[133] TUKIAINEN, T., PIRINEN, M., SARIN, A. P., LADENVALL, C., KETTUNEN, J., LEHTIMAKI, T., LOKKI, M. L., PEROLA, M., SINISALO, J., VLACHOPOULOU, E., ERIKSSON, J. G., GROOP, L., JULA, A., JARVELIN, M. R., RAITAKARI, O. T., SALOMAA, V., AND RIPATTI, S. Chromosome X-wide association study identifies Loci for fasting insulin and height and evidence for incomplete dosage compensation. *PLoS Genet 10*, 2 (2014), e1004127.

[134] UK IBD GENETICS CONSORTIUM, BARRETT, J. C., LEE, J. C., LEES, C. W., PRESCOTT, N. J., ANDERSON, C. A., PHILLIPS, A., WESLEY, E., PARNELL, K., ZHANG, H., DRUMMOND, H., NIMMO, E. R., MASSEY, D., BLASZCZYK, K., ELLIOTT, T., COTTERILL, L., DALLAL, H., LOBO, A. J., MOWAT, C., SANDERSON, J. D., JEWELL, D. P., NEWMAN, W. G., EDWARDS, C., AHMAD, T., MANSFIELD, J. C., SATSANGI, J., PARKES, M., MATHEW, C. G., WELLCOME TRUST CASE CONTROL, C., DONNELLY, P., PELTONEN, L., BLACKWELL, J. M., BRAMON, E., BROWN, M. A., CASAS, J. P., CORVIN, A., CRADDOCK, N., DELOUKAS, P., DUNCANSON, A., JANKOWSKI, J., MARKUS, H. S., MATHEW, C. G., MCCARTHY, M. I., PALMER, C. N., PLOMIN, R., RAUTANEN, A., SAWCER, S. J., SAMANI, N., TREMBATH, R. C., VISWANATHAN, A. C., WOOD, N., SPENCER, C. C., BARRETT, J. C., BELLENGUEZ, C., DAVISON, D., FREEMAN, C., STRANGE, A., DONNELLY, P., LANGFORD, C., HUNT, S. E., EDKINS, S., GWILLIAM, R., BLACKBURN, H., BUMPSTEAD, S. J., DRONOV, S., GILLMAN, M., GRAY, E., HAMMOND, N., JAYAKUMAR, A., MCCANN, O. T., LIDDLE, J., PEREZ, M. L., POTTER, S. C., RAVINDRARAJAH, R., RICKETTS, M., WALLER, M., WESTON, P., WIDAA, S., WHITTAKER, P., DELOUKAS, P., PELTONEN, L., MATHEW, C. G., BLACKWELL, J. M., BROWN, M. A., CORVIN, A., MCCARTHY, M. I., SPENCER, C. C., ATTWOOD, A. P., STEPHENS, J., SAMBROOK, J., OUWEHAND, W. H., MCARDLE, W. L., RING, S. M., AND STRACHAN, D. P. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat Genet 41*, 12 (2009), 1330–1334.

[135] VAN RAALTE, D. H., LI, M., PRITCHARD, P. H., AND WASAN, K. M. Peroxisome proliferator-activated receptor (PPAR)-alpha: a pharmacological target with a promising future. *Pharm Res 21*, 9 (2004), 1531–1538.

[136] VOSKUHL, R. Sex differences in autoimmune diseases. *Biol Sex Differ 2*, 1 (2011), 1.

[137] WAKELEY, J., AND HEY, J. Estimating ancestral population parameters. *Genetics 145*, 3 (1997), 847–855.

[138] WELTER, D., MACARTHUR, J., MORALES, J., BURDETT, T., HALL, P., JUNKINS, H., KLEMM, A., FLICEK, P., MANOLIO, T., HINDORFF, L., AND PARKINSON, H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res 42*, Database issue (2014), D1001–D1006.

[139] WILLER, C. J., LI, Y., AND ABECASIS, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics 26*,

17 (2010), 2190–2191.

[140] WILLIAMS, O. D. The atherosclerosis risk in communities (aric) study - design and objectives. *American Journal of Epidemiology 129*, 4 (1989), 687–702.

[141] WILSON, M. A., AND MAKOVA, K. D. Genomic analyses of sex chromosome evolution. *Annu Rev Genomics Hum Genet 10* (2009), 333–354.

[142] WISE, A. L., GYI, L., AND MANOLIO, T. A. eXclusion: toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet 92*, 5 (2013), 643–647.

[143] WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M., AND LIN, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet 89*, 1 (2011), 82–93.

[144] ZAITLEN, N., PASANIUC, B., PATTERSON, N., POLLACK, S., VOIGHT, B., GROOP, L., ALTSHULER, D., HENDERSON, B. E., KOLONEL, L. N., LE MARCHAND, L., WATERS, K., HAIMAN, C. A., STRANGER, B. E., DERMITZAKIS, E. T., KRAFT, P., AND PRICE, A. L. Analysis of case-control association studies with known risk variants. *Bioinformatics 28*, 13 (2012), 1729–1737.

[145] ZAYKIN, D. V., ZHIVOTOVSKY, L. A., WESTFALL, P. H., AND WEIR, B. S. Truncated product method for combining *P*-values. *Genet Epidemiol 22*, 2 (2002), 170–185.

[146] ZHANG, J., WHEELER, D. A., YAKUB, I., WEI, S., SOOD, R., ROWE, W., LIU, P. P., GIBBS, R. A., AND BUETOW, K. H. SNPdetector: a software tool for sensitive and accurate SNP detection. *PLoS Comput Biol 1*, 5 (2005), e53.

[147] ZHENG, G., JOO, J., ZHANG, C., AND GELLER, N. L. Testing association for markers on the X chromosome. *Genet Epidemiol 31*, 8 (2007), 834–843.